A Computational and Experimental Study of the Structure of the C-terminal Domain of the Wildtype and Mutant FOXL1 Protein

By

© Jessica Elizabeth Besaw

A thesis submitted to the

School of Graduate Studies

in partial fulfillment of the requirements for the degree of

Masters of Science

in

The Department of Chemistry

Faculty of Science

Memorial University of Newfoundland

August, 2015

St. John's, Newfoundland and Labrador

Abstract

Forkhead box (FOX) proteins are transcription factors that play a significant role during embryonic development and throughout adulthood with functions in regulating cell differentiation, proliferation, and apoptosis. Consequently, mutated or unregulated FOX proteins have been linked with numerous human genetic diseases. Recently, a deletion of five residues (GIPFL) in the C-terminal domain of the FOXL1 protein has been linked to another human genetic disease. This manuscript presents attempts to determine the three-dimensional structure of the C-terminal domain of FOXL1 protein, both with the mutation ($FOXL1_{MUT}$) and without ($FOXL1_{CTERM}$), in order to uncover the structural features that prevent the proper functioning of the mutant. This structural information was obtained using both computational and experimental methods. Computationally, coarse-grain molecular dynamic (MD) simulations were performed using replica exchange MD to provide a prediction of the folded structure. Experimentally, the C-terminal domain of FOXL1 and its mutant was expressed, enriched, and then structurally characterized using circular dichroism.

Bioinformatics analysis of FOXL1 revealed that the mutation occurred in the most ordered and evolutionary-conserved portion of the C-terminal domain of FOXL1, suggesting that this mutation could severely affect the structure and function of $FOXL1_{CTERM.}$ This is in agreement with the replica exchange molecular dynamics simulations results, which predicted that $FOXL1_{CTERM}$ was more folded and structured than $FOXL1_{MUT}$. Moreover, the simulations showed that the mutation in the $FOXL1_{MUT}$ system disrupted its structure and hydrophobic core, causing the mutant to have an increased amount of randomly coiled structure. The computational results are supported by preliminary experimental data. Circular dichroism results indicated that $FOXL1_{CTERM}$ has a predominantly helical structure while $FOXL1_{MUT}$ was partially helical with some randomly coiled regions.

Acknowledgements

There are numerous individuals I would like to recognize for their guidance and support during this masters project. First, I would like to thank my supervisors Dr. Christopher N. Rowley and Dr. Valerie Booth for their advice, assistance, and teachings, which were invaluable in the completion of this work. Thank you both for the opportunity to work on this collaborative project, which allowed me to gain skills in both computational chemistry and experimental protein synthesis. In particular, I want to thank Christopher Rowley for leading by example on how to be an amazing and successful teacher, researcher, and supervisor. I admire your clear and organized teaching style, as well as your involvement on a number of diverse research topics. I also want to thank you for being incredibly speedy with all replies including answering very late night emails as well as for fixing multiple computer crashes. Although you still remain mysterious on what your middle initial "N" stands for, your support, kindness, and incredible supervising skills are clear and evident. I also wanted to specifically thank Dr. Valerie Booth for offering the mini-course on "How to Be a Better Scientist." This course opened my eyes to many aspects of research and provided me with numerous strategies that I will definitely apply when pursuing a Ph.D. in chemistry. I am truly lucky to have two amazing mentors who would go above and beyond the call of duty as supervisors. In addition, thank you to my colleagues Saleh Riahi, Ernest Williams, Jennifer Smith, Kari Gaalswyk, and Archita Adluri for sitting through many practice presentations and providing me with feedback. I enjoyed talking to every single one of you throughout the day. I want to acknowledge the crucial information concerning the mutant FOXL1 gene and the linked human genetic disease provided by Dr. Terry Young, without whom this project would not exist. In addition, I want to thank my supervisory committee member, Dr. Erika Merschrod, for donating her time to ensure my research was on track and for providing excellent suggestions for improving my thesis.

I would like to thank the Department of Chemistry and the Department of Biochemistry for the use of their facilities and resources. I would like to thank the Natural Science and Engineering Research Council for financial support provided through a Julie-Payette NSERC scholarship during my first year of research. I would also like to thank the School of Graduate studies for financial

support provided through the A. G. Hatcher Memorial Scholarship during the second year of research. Finally, I gratefully acknowledge the support of SciNet and Compute Canada for providing computer time.

I would like to extend a special thanks to my friends, especially my BESTEST BEST FRIEND EVER, Ashley Quirke, for always being up for a phone call, providing me with encouragement on really difficult days, and making me laugh out loud. I want to thank her for always lending a helping hand or a listening ear when needed. I will always remember Punta Cana as one of the best trips that we ever took next to Mexico, Florida, Gros Morne, Kingston, and Toronto. Hopefully we will be travelling buddies forever! I also want to thank my running buddy Matt Hunt for motivating me to exercise! Thank you for running Cape to Cabot with me (twice!) even though it was brutal, cold, and miserable outside. And moreover, thanks for all the fun and interesting conversations during those long half marathon training session. I also want to thank my boyfriend Colin Ash for being my study buddy during many late nights at MUN. I appreciate the encouragement you gave me to compete for the MUN cross-country team. I loved sharing numerous laughs and fun times, even though you are an incessant troll! Most importantly, thanks for proof-reading this thesis!

I want to thank my mom for supporting me through late night research. Mom, thank you for making me lunch and supper every single work day. Your help provided me with more free time that I could put into my masters project while not starving in the process! Also, thank you for checking in on me during the very late hours of the night that I would spend at MUN. I know that if anything happened, you would know right away. Also, thank you Dad for providing me with everything that I need including a loving home, food on the table, and the clothes on my back. I really enjoyed our discussions on the economy, which helped me make better financial choices. Dad, thanks for the Toyota Rav 4! Now, I can drive towards my future. The continued support and encouragement from family will always be greatly appreciated.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BCIP	: 5-bromo-4-chloro-3-indolyl phosphate
BLASTP	: Basic Local Alignment Search Tool for Proteins
CAPS	: 3-cyclohexylamino-1 propane sulfonic acid
CD	: circular dichroism
CG	: coarse grain
CGW	: coarse grain water
CHAPS	: 3-(3-(cholampidopropyl) dimethylammonio)-1-propanesulfonate
DE52	: diethylaminoethyl, anion exchange pre-swollen whatman cellulose resin
DNA	: deoxyribonucleic acid
E-coli	: *Escherichia coli*
EDTA	: ethylenediaminetetraacetic acid
EM	: electron microscopy
FOXL1	: forkhead box L1
$FOXL1_{CTERM}$	: C-terminal forkhead box L1 protein
$FOXL1_{MUT}$	: mutant C-terminal forkhead box L1
FT	: flow-through
GIPFL	: glycine-isoleucine-proline-phenylalanine-leucine
IDP	: intrinsically disordered protein
IDR	: intrinsically disordered region
IMAC	: immobilized metal ion affinity chromatography
Imdzl	: imidazole
IPTG	: isopropyl-D-β thiogalactopyranoside
L.1	: lane 1 on a gel or western blot
MD	: molecular dynamics
MWCO	: molecular weight cut-off

NBT	: nitro blue tetrazolium chloride
NMR	: nuclear magnetic resonance
NOESY	: nuclear Overhauser effect spectroscopy
PACE	: protein in atomic details coupled with coarse-grained environment
PAGE	: polyacrylamide gel electrophoresis
PDB	: protein data bank
PMSF	: phenylmethylsulfonyl fluoride
PVDF	: polyvinylidene fluoride
REMD	: replica exchange molecular dynamics
SDS-PAGE	: sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SL	: sample loaded
SN	: Staphylococcus aureus nuclease
TBS	: tris-buffered saline
TE	: Tris and EDTA
TOCSY	: total correlation spectroscopy
TTBS	: tween-tris buffered saline
2xYT	: 5 g/L NaCl, 10 g/L yeast and 16 g/L tryptone
6His	: His-His-His-His-His-His

Chapter 1

Introduction

1.1. Protein Structure and Function

Proteins are large biomolecules which serve many vital roles in all living organisms, including catalyzing reactions, replicating DNA, and transporting molecules from one location to another.[1] A protein is composed of one or more chains of amino acid residues that fold into complex three-dimensional structures. The amino acid sequence of a protein ultimately determines its three-dimensional structure.[1] Since the proper functioning of proteins is greatly dependent on their structure, a mutation in the amino acid sequence can cause the protein to fold incorrectly and lead to severe human diseases.[2–4]

Determining structural information about proteins and their mutants can yield insight into their function or misfunction.[1,3] The configuration of a protein can be classified by its primary, secondary, and tertiary structure, as shown in Figure 1.1. The primary structure is the sequence of amino acids that constitute the protein. The secondary structure is the local shape of a portion of the protein brought about though hydrogen bonding between the amide groups of the protein backbone. The major secondary structure motifs include α-helices (coil-shape) and β-sheets. The

tertiary structure is the overall three-dimensional structure that results from protein folding such that contacts are formed by the secondary structural elements. Additionally, if a protein is composed of multiple folded polypeptide chains, it can further be classified by its quaternary structure. The quaternary structure is the number and arrangement of the polypeptide subunits with respect to each other.

Proteins, such as those investigated in this thesis (see Section 3.1.1 below), do not necessarily need to have a unique, ordered structure to perform a function. There are numerous intrinsically disordered proteins (IDPs) that lack well-defined three-dimensional shape under physiological conditions and still perform biologically relevant functions.[5,6] These IDPs can be fully disordered or possess local regions of disorder referred to as intrinsically disordered regions (IDRs).[6]

In general, IDPs and IDRs are often characterized by low-complexity sequences, which have little diversity in the amino acid composition and are generally made up of just a few different amino acids. In particular, IDPs and IDRs are mainly composed of polar and charged amino acids (e.g. arginine (R), lysine (K), glutamate (E), proline (P), and serine (S)), while having low levels of bulky hydrophobic amino acids (e.g. tryptophan (W), tyrosine (Y), isoleucine (I), and valine (V)).[5] IDPs generally carry out their function by binding to a target molecule, often another protein, and fold into a defined structure upon binding.[5]

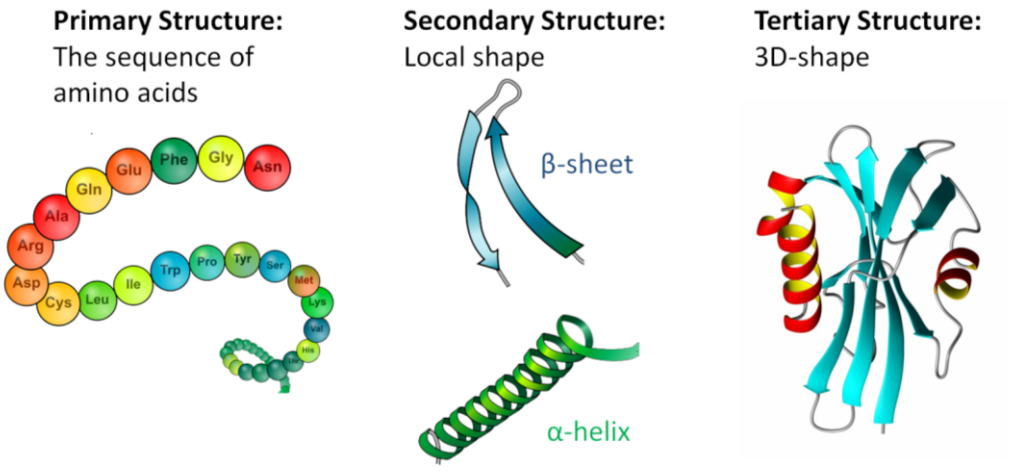


Figure 1.1: Main protein structure levels. Figure adapted from "Main protein structure levels" by Mariana Ruiz Villarreal which was released into the public domain in 2008, and "Ribbon diagram of a protein" by David E. Volk which was released into the public domain in 2007.

1.2. FOXL1 Protein

The forkhead box (FOX) class of proteins exhibit important structure-function relationships that are essential to human health.[3,4] FOX proteins have a structured N-terminal domain, known as the "forkhead" domain, that binds to DNA in order to regulate transcription.[3,4,7] Through transcriptional regulation, FOX proteins influence a diverse range of cellular and developmental processes.[4,8–13] Considering their crucial functions in humans, it is not surprising that mutated or unregulated FOX proteins have caused human genetic diseases that can surface during embryonic development and throughout the lifetime of an adult.[3] As of 2003, mutations in 11 FOX proteins have been linked to human genetic diseases.[3] Premature ovarian failure, mental retardation, and severe immune defects are just a few of the severe health problems linked with mutations in the FOXO3a, FOXP1, and FOXN1 proteins, respectively. Recently, a mutation in FOXL1 protein has been shown by Dr. Terry Young[†] and coworkers to be associated with a human genetic disease (unpublished). Thus, the structure of FOXL1 and the effect of this mutation on its structure is of high interest.

The tertiary structure of the N-terminal DNA binding domain is similar for all forkhead box proteins, including FOXL1. The structure of the forkhead domain was determined experimentally by nuclear magnetic resonance for FoxC2[‡].[14] This forkhead domain is composed of approximately 80 to 100 amino acids[3,15] which forms three α-helices, three β-strands, and two wing-shaped loops, as seen for FoxC2 in Figure 1.2. A loop is a section of amino acids that does not have fixed internal hydrogen bonding, which is often seen connecting regular secondary structural elements. The forkhead motif has been referred to as a winged helix in the literature because the loops have a butterfly-like appearance.[16] The alignment of the N-terminal structure with the amino acid sequence of FOXL1[14] is shown in Figure 1.3.

In contrast to the N-terminal domain, the C-terminal region of forkhead proteins are highly divergent, leading to a variety of classes of forkhead box proteins that each have a different C-terminal structure and function.[15] In fact, more than one hundred members of the forkhead gene

[†] Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada

[‡] Forkhead box nomenclature contains all uppercase letters for humans (FOXL1), only the first letter capitalized for mice (Foxl1), and the first and subclass letter capitalized for all other chordates (FoxC2).[15]

family have been identified. In particular, humans have 50 FOX genes which code for 50 FOX proteins, one of which is FOXL1.[3] Due to the overwhelming number of forkhead proteins, the tertiary structure of the C-terminal region of most of these proteins have not been characterized, including FOXL1.

FOXL1 is a 345 amino acid protein with three known functions.[16–19] First, it is a transcription factor required for proper proliferation and differentiation in the gastrointestinal epithelium.[10–12] Secondly, in the Wnt/β-catenin pathway, the FOXL1 protein is involved in the inhibition of Wnt signalling.[12] Finally, in the Sonic-Hedgehog signalling pathway, the FOXL1 protein is involved in mediation of endoderm-mesoderm signals.[8]

The 345 amino acid sequence of FOXL1 is detailed in Figure 1.3. The well-studied N-terminal forkhead domain spans residue 48-139. This domain is composed of three α-helices (54-63; 72-83; 92-99), three β-strands (68-70; 107-112; 120-125), and two wing-shaped loops (113-119; 126-140) structural elements. The mutation in FOXL1 under investigation here corresponds to the deletion of five amino acids involving residues 326-330, which is highlighted in red. The deleted sequence consists of a glycine (G), isoleucine (I), proline (P), phenylalanine (F), and leucine (L), which this thesis refers to as the GIPFL deletion.

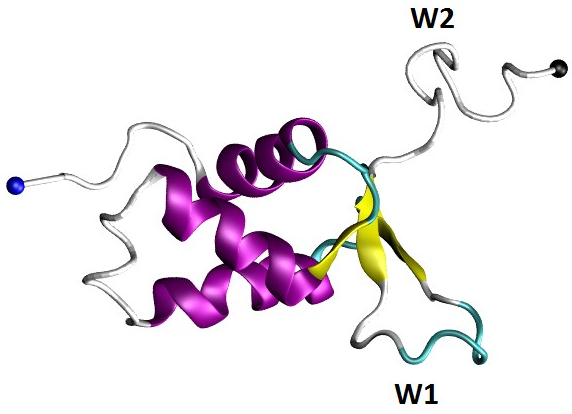


Figure 1.2: The structure of the forkhead domain of FoxC2 (UniProt ID Q99958) has three α-helices (purple spirals), three β-strands (yellow arrows), and two wing-shaped loops (labelled W1 and W2). The blue and black spheres marks the N- and C-terminus, respectively. The coordinate file was downloaded from the PDB website[§] (PDB ID 1D5V) and rendered using VMD 1.9.1.

[§] http://www.rcsb.org/pdb/home/home.do

The GIPFL deletion occurs in the C-terminal region of FOXL1, which has not yet been structurally characterized. Unfortunately, it is difficult to study the full 345 amino acid FOXL1 protein both computationally and experimentally. A large protein system is computationally demanding because it requires long simulation times to sample the conformational space. It is also experimentally challenging to elucidate the structures of large proteins using spectroscopic techniques like NMR because they often contain a large quantity of signals in the spectrum. Thus, this thesis will investigate only the C-terminal domain of FOXL1. A domain is a section of a protein that can maintain its structure and function independently of the rest of the protein chain. Here, the C-terminal domain is defined as the sixty-nine most C-terminal residues of FOXL1, which we refer to as $FOXL1_{CTERM}$. The details of how the C-terminal domain was determined can be found in Section 3.1.1. The mutant of C-terminal FOXL1 protein, which we refer to as $FOXL1_{MUT}$, is defined here as the same residues as $FOXL1_{CTERM}$ with the GIPFL deletion.

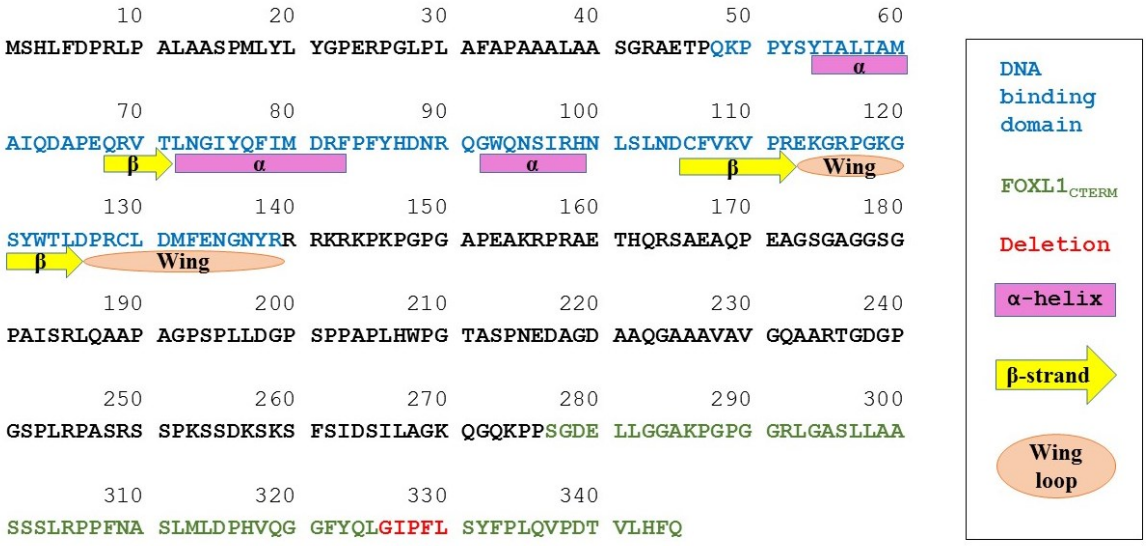


Figure 1.3: The FOXL1 protein sequence is composed of 345 amino acids. FOXL1 has a characterized N-terminal DNA binding domain spanning residue 48-139, which is made up of three α-helices (54-63; 72-83; 92-99), three β-sheets (68-70; 107-112; 120-125), and two wing-shaped loops (113-119; 126-140) structural elements. The mutant FOXL1 has five deleted amino acids residues 326-330 (GIPFL), which are highlighted in red. The structures studied in this thesis are $FOXL1_{CTERM}$ (highlighted with green and red) and $FOXL1_{MUT}$ (highlighted in green).

The goal of this research is to determine the structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ in order to gain insight into the structural differences that causes the loss of function in the FOXL1 mutant. In this thesis, both computational and experimental methods are used in an attempt to determine the structure $FOXL1_{CTERM}$ and $FOXL1_{MUT}$.

1.3. Methods to Predict Protein Structure

1.3.1. Introduction

Several popular methods that are employed to predict protein structure include bioinformatics,[6,20,21] homology modelling,[22,23] and molecular simulations.[24–27] These methods all require the amino acid sequence of the protein in order to make predictions of the protein structure. The full FOXL1 sequence can be found on the UniProt database using the identification Q12952.[**] The information acquired from these methods can be very valuable in predicting protein disorder, secondary structure, sequence conservation, and even tertiary structure. The method of choice largely depends on the protein studied and whether or not structural information is available on other proteins with similar amino acid sequences.

1.3.2. Bioinformatics

The first step in investigating protein structure should involve employing bioinformatics to predict the properties and structure of the protein. Structural bioinformatics is the field of computational biology that infers the properties and structure of macromolecules based on generalizations made from experimentally and computationally solved structures.

Bioinformatics tools are often freely available online and provide a quick, cheap, and easy method to gain insight into the properties of a protein. In the following sections, several bioinformatics tools are discussed, which are later utilized on FOXL1 to determine intrinsically disordered regions,[6] identify domains, predict secondary structure,[28] and perform sequence alignment.[20]

[**] http://www.uniprot.org/uniprot/Q12952

1.3.2.1. Protein Disorder

A protein disorder predictor computationally deduces the ordered and intrinsically disordered regions of a protein.[5] Intrinsically disordered regions (IDRs) are partial regions of proteins which lack stable and well defined three-dimensional structure.[6]

There are numerous protein disorder software packages freely available. These predictors differ in the input required as well as the predictive approach used to assign disorder. The predictor may require as input either amino acid sequence, sequence complexity, amino acid composition, position specific scoring matrices, predicted secondary structure, or predicted accessible surface area.[6] These methods use a variety of algorithms to estimate the levels of protein disorder, including artificial neural networks, super vector machines, Baysesian methods, decision tree bases methods, template sequences, or charge-hydropathy plots.[6]

In this thesis, PONDR-FIT was selected to predict the structured and unstructured regions of FOXL1 because of the numerous advantages associated with using this disorder program.[6] First, PONDR-FIT only requires the amino acid sequence as input, which is readily available for FOXL1. In addition, PONDR-FIT is a meta-predictor, meaning it combines six individual disorder predictors (PONDR-VLXT, PONDR-VSL2, PONDR-VL3, FoldIndex, IUPred, and TopIDP) to create an eight-fold cross-validation disorder predictor. The benefit of combining these six individual predictors is that each uses a different predictive approach, and thus helps emphasize different features of the sequence. Xue *et. al.,* the developers of PONDR-FIT, showed that combining these six predicators improves the accuracy between 3-20% compared to the individual intrinsic disorder predictors.[6] They further demonstrated that PONDR-FIT was more accurate than each of the individual predictors across all datasets studied including fully ordered, fully disordered, and partially disordered datasets.[6] This is important for this study because FOXL1 has a structured N-terminal DNA binding domain, but we do not know the state of the C-terminal region. Overall, PONDR-FIT has an estimated 85% prediction accuracy based on sensitivity for the disordered residues.[6]

However, like all protein predictors, PONDR-FIT has limitations. First, disorder predictors are not very accurate along the boundary between the structured and intrinsically disordered regions.[6] Another drawback is that there is poor accuracy for disordered regions of ten residues or less.[6]

The importance of protein disorder predictions is that they can provide insight into the existence of several domains within a protein. This has applications in reducing the size of the system studied. A disordered analysis was performed on FOXL1 to determine if the mutation occurs in a distinct domain of the protein.

1.3.2.2. Sequence Alignment

In bioinformatics, a sequence alignment of a protein is a way of arranging the amino acid sequence of two or more proteins in order to identify regions of similarity.[29] The results of a sequence alignment can provide structural and sequence conservation information. This arises because proteins sharing similar sequences often have similar structural, functional, or evolutionary relationships. Moreover, the degree of similarity between proteins can be interpreted as a rough measure of how well a particular region is evolutionarily conserved.[30] A more conserved region of a protein suggests that the region is structurally or functionally important. In reference to the study of FOXL1, a sequence alignment is important to find similar proteins with already known structures for homology modelling, as well as to identify conserved amino acids in the C-terminal domain.

The Basic Local Alignment Search Tool for Proteins (BLASTP)[20,23,29] is a free, online, heuristic sequence alignment program that finds regions of local similarity between the query protein and those in the protein data bank (PDB). To perform a sequence alignment, the BLASTP algorithm (1) compiles a list of "high scoring", consecutive amino acid "words" from the query protein, (2) scans the database to find matches, and (3) extends the match to find the maximum segment pair alignment.[20] To quantify the alignment, BLASTP employs a PAM-120 matrix[31,32] of similarity scores to score all possible pairs of residues. In this matrix, identical pairs of residues have the highest positive score, conservative replacements (where the chemical properties of the amino acid stays the same) also have positive scores, while unlikely replacements have negative scores.

There are numerous advantages and several limitations of using BLASTP. First of all, the BLASTP algorithm is simple, robust, fast, and free.[20] As well, BLASTP is of comparable sensitivity and an order of magnitude faster than other current heuristic sequence alignment approaches.[20] In addition, by nature of the local similarity algorithm, BLASTP can find distantly related proteins, such as those with a similar active site. However, since the local similarity algorithm depends on

aligning consecutive amino acids words, it can miss biologically similar sequences that have a global alignment but do not have many words in common.[20] BLASTP may also overlook related proteins if the input sequence has a uniform pattern of conservation with few high scoring words. Finally, although BLASTP can find biologically significant relationships, a search can also result in chance similarities and it is left to the researcher to remove these irrelevant results.[20]

A sequence alignment is a valuable tool to deduce structural and sequence conservation information. In reference to the study of FOXL1, a sequence alignment is important to identify conserved amino acids in the C-terminal domain, especially concerning the GIPFL residues. Furthermore, if similar proteins with already known structures are found, then structural information through homology modelling could be acquired.

1.3.2.3. Secondary Structure Prediction

Several one-dimensional predictors have been created that use the protein sequence to predict its secondary structure.[28,33] PredictProtein is an online server that encompasses numerous programs that perform sequence alignment, predict structure, and infer functions of proteins.[33] One utility found under the umbrella of the PredictProtein server is PROFsec: a one-dimensional predictor of secondary structure.[28] To construct a structure prediction, PROFsec takes the amino acid sequence as input, and then employs evolutionary information generated by a multiple sequence alignment and feeds this alignment into a neural network system.[28] After multi-level processing to obtain one-dimensional information and filtering to remove drastic, unrealistic predictions, PROFsec outputs a probable secondary structure for each amino acid in the protein. Each amino acid is labelled as one of three secondary structure states encompassing helix (H = α, π, or 3_{10}), extended strand (E = parallel or anti-parallel β-sheet), and loop (L).[28]

Although several secondary structure predictors exist, PROFsec was chosen because it has a high prediction accuracy of $72\% \pm 9\%$.[33] However, the PROFsec program was trained on globular, water soluble proteins, and thus the projected accuracy is likely only valid for proteins of the same type.[28] Another limitation of PROFsec is that the accuracy of the secondary structure prediction is highly dependent on the accuracy of the sequence alignment and the number of diverse alignments made.[28] In fact, poor alignments have been shown to give poor secondary structure predictions, even for homologous proteins.[28] In addition, rare folds are rejected by the

PROFsec algorithm, which leads to incorrect predictions for proteins that do possess these unusual structural features. One final limitation of PROFsec is that the predictor is less reliable along the ends (caps) of helices and strands as compared to the core.[28] Thus, a predicted secondary structure segment can be longer or shorter than the true structure.

Secondary structure predictions are valuable tools that can make accurate predictions of protein secondary structure from the primary structure. In reference to the study of FOXL1, a structure prediction gives us a first look at the structural blueprint and can perhaps be used as a starting point for computer simulations.

Although bioinformatics provides a quick and easy way to deduce information about protein structure, it is not a replacement for an experimentally determined structure that provides the tertiary structure that is typically responsible for protein function.

1.3.3. Homology Modeling

Homology modeling is a technique employed to model a "target" protein of unknown structure by comparing it to a homologous "template" protein of known 3D structure.[22,23,30,34,35] This method assumes that proteins with similar amino acid sequences have similar structures.[22,30] This assumption is well-supported for proteins found in nature: evolutionarily related proteins with 50% sequence identity usually have near identical tertiary structures.[30] However, this assumption does not extend to artificially similar proteins, or pseudohomologs, that are synthetically designed to have similar sequences with different structures.[30]

The homology modeling process involves (1) identifying the best template protein, (2) aligning the target to the template sequence, (3) building a three-dimensional model of the target protein from the alignment, and finally (4) evaluating the reliability of the model.[23,30] There are numerous homology modeling programs available including Modeller, SegMod/ENCAD, SWISS-MODEL, 3D-JIGSAW, nest, and Builder.[22] In a benchmark study of these programs, it has been shown that no single program outperforms the others in all tests, including reliability, speed, structural accuracy, and physiochemical correctness.[22] However, Modeller is currently the most popular because it is fast and free for academic use.[22]

In the absence of an experimentally known structure, homology modelling can be a faster approach to elucidating protein structure.[23] This technique can provide insight into structural and evolutionary features of a protein, and is useful in drug design efforts.[22] Unfortunately, homology modelling is only practical if a suitable template protein exists. The accuracy of the generated structure greatly depends on the choice of template protein and the quality of the target-template sequence alignment.[22,23] Another problem is that it can be difficult to model the effects of mutations, because the mutant protein will often have the same template protein as its naturally occurring form.[22]

Homology modeling can provide a good starting point for predicting the tertiary structure of a protein. In fact, the structure generated from homology modelling can be further used in molecular dynamics simulations to make inferences about the kinetics and dynamics of the protein.

1.3.4. Computational

Theoretical chemistry combines the laws of physics and chemistry with mathematical tools in order to rationalize chemical phenomena.[36] This field has opened the doors to describe, explain, and predict the chemistry of molecules. Computational chemistry integrates theoretical chemistry with computer science in order to study, model, and simulate chemical systems.[37] As computing power increases and efficient algorithms are designed, computational chemistry methods can deal with large biomolecular systems with better accuracy. There are numerous computational chemistry tools that range from highly accurate methods, which are derived from quantum mechanics such as *ab initio* and density functional methods,[36] to highly approximate methods, which are based on classical descriptions of forces such as molecular mechanics.[38–40] In order to determine the structure of an unknown protein, the less computationally expensive molecular mechanics approximations are generally employed and coupled with classical molecular dynamics simulations to fold a protein into its native structure.[38]

1.3.4.1. Molecular Dynamic Simulations

Protein folding by molecular dynamics simulations explores the process by which a protein changes its conformational state during a trajectory to obtain a more stable, lower energy

structure.[41] To fold a protein, computer simulations must first model the physical forces in chemical systems. The major factors that contribute to protein folding include hydrogen bonding, van der Waals interactions, backbone angle preferences, electrostatic interactions, hydrophobic interactions, and chain entropy.[41] These physical forces can be described classically through a force field: a mathematical model used to approximate the potential energy of a system.[24,25] In a computer simulation, a protein is placed in an initial unfolded configuration. By using a force field to determine the energy of the system (Equation 1.1) and employing dynamical laws of motion to propagate the position of the atoms through time by a series of times steps (Δt) (i.e. the trajectory shown in Equation 1.2 and Equation 1.3), changes in the conformation of a protein can be modelled during the course of a simulation.[41]

$$U(\boldsymbol{r}) = \sum_{\text{bonds}} \frac{1}{2} k_{bond}(r - r_e)^2 + \sum_{\text{angles}} \frac{1}{2} k_\Theta(\Theta - \Theta_e)^2 + \sum_{\text{dihedrals}} \frac{1}{2} V_\phi(1 + \cos(n\phi - \delta)) + \sum_{\text{pairs}} 4\,\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \sum_{\text{pairs}} \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} \quad \textbf{(1.1)}$$

$$\boldsymbol{F} = -\nabla U \quad \textbf{(1.2)}$$

$$\boldsymbol{r_i}(t + \Delta t) = 2\boldsymbol{r_i}(\mathrm{t}) - \boldsymbol{r_i}(t - \Delta t) + \frac{\boldsymbol{F_i}(t)}{m_i}\Delta t^2 \quad \textbf{(1.3)}$$

An important consideration prior to running a computer simulation is the resolution of the system and the chosen force field.[42] In general, a particular force field is tied to a specific scheme for atomic representations. The force field chosen greatly affects the accuracy of the potential energy calculations and the molecular dynamics simulation, so this decision should be given significant consideration.

The highest resolution models in molecular mechanics are the all-atom models where each atom is represented as an interaction point.[36,37] The potential energy, $U(r)$, for an all-atom force field is the total energy of all the covalent interactions including vibrational, bending, and torsional energy, as well as non-covalent interactions such as the Lennard-Jones and electrostatic potential.[43] The potential energy can be calculated using Equation 1.1, where the variables highlighted in red are the parameters of the force field. The vibration term requires parameters

for the force constant (k_{bond}) and equilibrium bond length (r_e) to calculate the vibrational energy at a bond length (r). The bending term requires parameters for the force constant (k_Θ) and equilibrium angle (Θ_e) to calculate the energy of bending at an angle (Θ). The torsional potential approximates the energy of twisting a bond based on the rotation potential (V_ϕ), the multiplicity (n), and the phase shift (δ) parameters. The non-covalent interaction between particles separated by a certain distance (r_{ij}) is modelled through a Lennard-Jones potential which requires parameters for the depth of the potential well (ε_{ij}) and the distance where the inter-particle potential is zero (σ_{ij}). Finally, the electrostatic interaction is modelled by a Coulomb potential that requires parameters for the charges on the particles (q_i and q_j), where ε_o is simply a constant for the permittivity of free space. All of these parameters must be pre-determined by fitting to experiment or *ab initio* calculations. Numerous all-atom force fields exist because there are a variety of different methods employed to obtain or estimate these parameters. Due to the many design choices inherent in parameterization, force fields yield considerable differences in predicted biophysical properties.[43]

After calculating the potential energy of a system using a particular force field, a computer simulation employs two mathematical equations in order to propagate the atoms through space and achieve a lower energy structure. First, the force on the system ($\boldsymbol{F}$) is calculated as the negative gradient of potential energy ($-\nabla U$), as seen in Equation 1.2. Then, the Verlet equation for the description of motion is used to propagate atoms, as shown in Equation 1.3. In the Verlet equation, for a particle with mass, m_i, the future position of a particle $(r_i(t+\Delta t))$ is based on its current $(r_i(\mathrm{t}))$ and previous $(r_i(t-\Delta t))$ positions, with a force $(F_i(t))$ acting on it over a time step Δt.

A variety of all-atom force fields for proteins have been developed including CHARMM (e.g. CHARMM27, CHARMM36), AMBER (e.g. ff03, ff96sb*, ff99sb-ildn-NMR), and OPLS-AA.[43] Another consideration for all-atom models is that proteins and biological systems are generally studied in a water solvent, and so the resolution of the solvent model must also be defined. The solvent model employed can be either explicit, such as TIP3P, SPC/E, and TIP4P-EW, or implicit such as GBSA.[43] Explicit solvent models represent the water molecules and salt ions as explicit particles at discrete spatial positions.[44] These models are advantageous because protein hydration and solvent entropic effects are described according to their fundamental physical interaction, but

disadvantageous because they require significant computational time.[41] On the other hand, implicit water models represent the solvent as a continuous medium with constant dielectric permittivity.[44] The benefit of implicit solvent is its fast equilibration with the downside that it is only an approximation to bulk water and there is no explicit water interacting with the protein. Overall, all-atom models most accurately represent the potential energy of system, but they require high simulation times and computational resources to simulate large protein systems.[24,25,42,45,46]

Coarse grain (CG) force fields enable faster molecular dynamic simulations than all-atom models.[42,45–50] In CG models, a group of atoms is treated as a virtual particle, or a "bead", in order to decrease the number of particles in a system and reduce the complexity of the force field, which leads to significant computational savings.[42,45–50] Several protein CG models have been introduced, including one-bead,[51] two-bead,[50] and multiple-beads per residue models,[49] as well as even much coarser models that have one bead representing multiple residues.[52] The most popular coarse grain model is MARTINI, which employs an intermediate CG environment where four heavy atoms (e.g. not hydrogen) are reduced to one site.[49] These coarse-grained models are often limited in their use for studying of folding, aggregation, large conformational changes of proteins, or conformational features of intrinsically disordered proteins. Due to the loss of atomistic information, many coarse-grained simulations of proteins rely on information from its native structure in order to apply constraints to maintain its structure. Many of these coarse grained models were unsuitable for the study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ because the native structure of this protein was unknown at the outset of this project.

A hybrid model has recently been developed by Wu and coworkers[24,25,45,53] which is capable of quickly and accurately folding proteins into their native structure without prior structural information. This hybrid model, known as PACE,[24,25,45,53] was employed to study the structure $FOXL1_{CTERM}$ and $FOXL1_{MUT}$.

1.3.4.2. PACE Force Field

In this research, $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ were represented using the Protein in Atomic details coupled with Coarse-grained Environment (PACE).[24,25,45] PACE represents protein and ions using the united atom (UA) model and represents water using a coarse grain environment developed

by Marrink and coworkers.[48] In the UA model, each heavy atom and its attached hydrogens are usually represented as a single bead,[45] as shown in Figure 1.4. However, hydrogen is explicitly represented in cases where hydrogen bonding can occur, such as the hydrogen in the backbone amide groups and in the side chains of Asn, Gln, Trp, and His. A coarse grain water bead is a sphere representing a cluster of four water molecules, as illustrated in Figure 1.5.

There were three major motivations for using the PACE force field as opposed to all-atom and other coarse grain models. First, simulations using PACE are approximately five to ten times faster than conventional all-atom simulations.[45] Secondly, PACE was shown to be capable of maintaining the native structure of protein in MD simulations, unlike many coarse grain models that require artificial constraints.[24,25,53] Thirdly, PACE was able to fold proteins from a random coil into their native structure. These proteins included a variety of structural motifs including α-helix, β-sheet, and mixed structures.[45,53] This is especially important in the structural study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ protein because no experimentally determined or homology model structure was available to serve as the starting structure.

The total energy of a protein system using the PACE force field is the sum of contributing interactions including bonded (E_{bond}), bending (E_{angle}), dihedral (E_{dihedral}), geometry (E_{improper}), rotamers (E_{φ,ψ,χ_1}), water ($E_{\mathrm{CGW-CGW}}$), hydration ($E_{\mathrm{CGW-UA}}$), van der Waals (E_{vdW}), polar (E_{polar}), and electrostatic (E_{ele}) interactions, as detailed in Equation 1.4. Han and co-workers, the developers of the PACE model, parameterized each of these terms. We refer the reader to several papers that describe the terms and parameterization in detail.[45,47,49,54]

$$E = E_{\mathrm{bond}} + E_{\mathrm{angle}} + E_{\mathrm{dihedral}} + E_{\mathrm{improper}} + E_{\varphi,\psi,\chi_1} + E_{\mathrm{CGW-CGW}} + E_{\mathrm{CGW-UA}} + E_{\mathrm{vdW}} + E_{\mathrm{polar}} + E_{\mathrm{ele}} \quad \textbf{(1.4)}$$

In the PACE force field, the bonding interaction, E_{bond}, is governed by a harmonic potential with a force constant of $K_{\mathrm{bond}} = 1.25 \times 10^5\ \mathrm{kJ\ nm^{-2}}$, as seen in Equation 1.5. The energy associated with bending, E_{angle}, is imposed with a harmonic potential with $K_{\mathrm{angle}} = 300\ \mathrm{kJ\ mol^{-1}\ rad^2}$, as shown in Equation 1.6.[24] The equilibrium bond lengths (r_0) and angles (Θ_0) were obtained by fitting these parameters to reproduce the structures obtained from quantum mechanical calculations of 15 small molecules.[47]

$$E_{\text{bond}} = \sum_{\text{b}} \frac{1}{2} K_{\text{bond,b}}(r_b - r_{0,b})^2 \quad \textbf{(1.5)}$$

$$E_{\text{angle}} = \sum_{\text{a}} \frac{1}{2} K_{\text{angle},a}(\Theta_a - \Theta_{0,a})^2 \quad \textbf{(1.6)}$$

The dihedral potential, E_{dihedral}, is made up a sum of cosine functions of the torsional angles and the Lennard-Jones potential for the interaction between sites separated by three bonds, as shown in Equation 1.7. The torsional potential contains terms for the rotational potential ($K_{\text{dih},d}$), the multiplicity (n_d), the dihedral angle (ζ_d), and the phase shift ($\zeta_{0,d}$). The Lennard-Jones potential includes terms for the depth of the potential well ($\varepsilon_{14,ij}$) and the distance where the inter-particle potential is zero ($\delta_{14,ij}$). These parameters were optimized by fitting the quantum mechanical torsional potential of 24 minima and 22 rotation barriers of simple molecules.[24] E_{improper} is a term used to keep planar geometries and chiral centers, which is calculated using Equation 1.8. This term is imposed using a harmonic potential on a dihedral (ξ_0) with $K_{\text{imp}} = 300 \text{ kJ mol}^{-1} \text{ rad}^2$. The energy of rotamers, E_{φ,ψ,χ_1}, for the backbone (ψ, φ) and sidechain (χ_1) is composed of a cosine function with different multiplicities (n) and a Lennard-Jones potential to model the short-range interaction between a side chain and its adjacent backbone for sites separated by more than three covalent bonds.[45] The parameters K_{dih}, ζ_{0_n}, $\varepsilon_{\text{short},ij}$, and $\sigma_{\text{short},ij}$, which are shown in Equation 1.9, were optimized through iterative equilibrium simulations to reproduce Ramachandran plots from a protein data bank coil library through aqueous simulation of dipeptides for each of the 20 amino acids.[24]

$$E_{\text{dihedral}} = \sum_{d} K_{\text{dih},d}\left[1 + \cos\left(n_d\zeta_d - \zeta_{0,d}\right)\right] + \sum_{1-4\text{pair}} 4\varepsilon_{14,ij}\left(\frac{\delta_{14,ij}^{12}}{r_{ij}^{12}} - \frac{\delta_{14,ij}^{6}}{r_{ij}^{6}}\right) \quad \textbf{(1.7)}$$

$$E_{\text{improper}} = \sum_{i} \frac{1}{2} K_{\text{imp},i}(\xi_0 - \xi_{0,i})^2 \quad \textbf{(1.8)}$$

$$E_{\varphi,\psi,\chi_1} = \sum_{d} \sum_{n_d=1}^{N_d} K_{\mathrm{dih}_{n_d,d}} \left[1 + \cos\left(n_d \zeta_d - \zeta_{0_{n_d,d}}\right)\right] + \sum_{\mathrm{short}} 4\varepsilon_{\mathrm{short},ij} \left(\frac{\sigma_{\mathrm{short},ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{\mathrm{short},ij}^{6}}{r_{ij}^{6}}\right) \quad \textbf{(1.9)}$$

A Lennard-Jones potential, shown in Equation 1.10, is employed to model the interaction between CG water ($E_{\mathrm{CGW-CGW}}$), between CG water and protein ($E_{\mathrm{CGW-UA}}$), and between non-polar protein sites (E_{vdW}). The parameters for the CG water interaction are $\varepsilon_{\mathrm{CGW-CGW}} = 5.0\ \mathrm{kJ\ mol^{-1}}$ and $\varepsilon_{\mathrm{CGW-CGW}} = 5.0\ \mathrm{kJ\ mol^{-1}}$.[45] The parameters for the interaction of protein and water were optimized by fitting experimental hydration free energies of 35 compounds.[45] The non-bonded interaction parameters were optimized to fit the density of liquid states and free energies of evaporations for eight organic compounds.[45]

$$E_{\mathrm{A-B}} = \sum_{i \neq j} 4\varepsilon_{A-B,ij} \left(\frac{\delta_{A-B,ij}^{12}}{r^{12}} - \frac{\delta_{AB,ij}^{6}}{r^{6}}\right) \quad \textbf{(1.10)}$$

The interaction between polar groups, E_{polar}, is calculated using Equation 1.11. The polar interaction parameters were optimized by fitting the potential of mean force (PMF) simulations to the OPLS-AA/L force field in explicit water. The electrostatic interaction is calculated using Equation 1.12, where q_i and q_j are atomistic charges, r_{ij} is the pair distance, ε_0 is the vacuum permittivity constant, and ε_r is the relative permittivity constant. The atomic charges were taken directly from the OPLS-AA force field.[45]

$$E_{\mathrm{polar}} = \sum_{|i-j|>2} \left[4\varepsilon_{\mathrm{attr}} \left(\frac{\delta_{Oi-NHj}^{12}}{r_{Oi-NHj}^{12}} - \frac{\delta_{Oi-NHj}^{6}}{r_{Oi-NHj}^{6}}\right) + 4\varepsilon_{\mathrm{rep}} \frac{\delta_{Oi-C\alpha j}^{12}}{r_{Oi-C\alpha j}^{12}} + 4\varepsilon_{\mathrm{rep}} \frac{\delta_{Oi-Cj-1}^{12}}{r_{Oi-Cj-1}^{12}} + 4\varepsilon_{\mathrm{rep}} \frac{\delta_{Ci-NHj}^{12}}{r_{Ci-NHj}^{12}}\right] \quad \textbf{(1.11)}$$

$$E_{\mathrm{ele}} = \sum_{i \neq j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \quad \textbf{(1.12)}$$

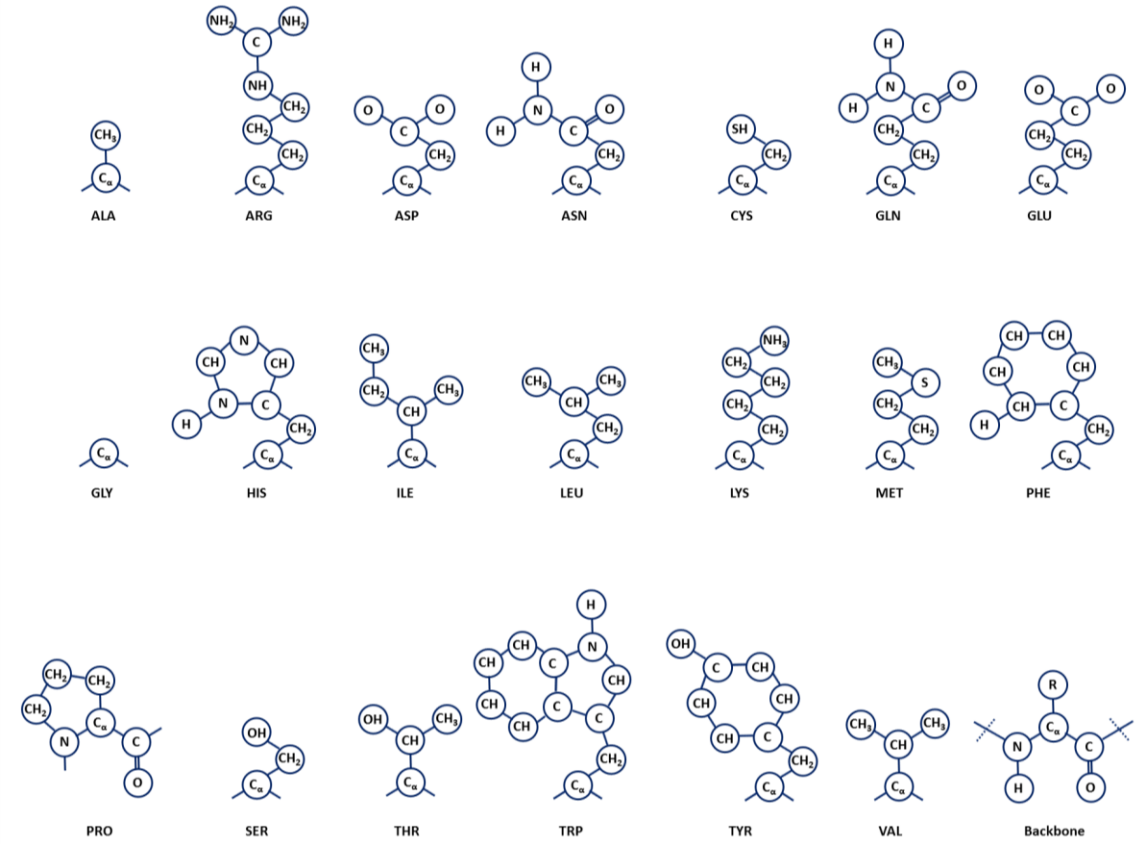


Figure 1.4: The United Atom representation of the amino acids employed in the PACE model. Each bead represents a single interaction site. Adapted with permission from Han *et. al.*[24] Copyright (2010) American Chemical Society.

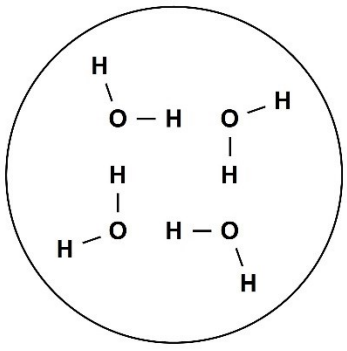


Figure 1.5: The coarse grain water used in the PACE model is a van der Waal's sphere representing a cluster of four water molecules, which was developed by Marrink *et. al.*[48]

1.3.4.3. Replica Exchange Molecular Dynamics

Replica exchange molecular dynamics (REMD) is a useful conformation sampling method for protein-folding, especially when attempting to fold a protein from a random coil to its native state.[27,34] In conventional molecular dynamics (MD) simulations, a protein can often become trapped in one of many local energy-minimum states. REMD resolves this problem by simulating a series of non-interacting replicas at several temperatures ranging from the desired temperature to a sufficiently high temperature that enables a replica to overcome energy barriers.[27] Throughout the simulation, the coordinates of temperature-neighbouring replicas can be exchanged at periodic intervals, where the probability ($P(i \rightarrow j)$) of switching coordinates is based on Metropolis-Hastings criterion shown in Equation 1.13. In this equation, E_i / E_j and T_i / T_j are the potential energy and temperature of the i^{th} and j^{th} replica, respectively, and k_b is the Boltzmann constant. The coordinate exchange of different temperature replicas enables greater conformational sampling, and increases the likelihood of determining the native state of a protein.[27,44]

$$P(i \rightarrow j) = \min\left(1\,,\ e^{(E_i - E_j)\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)}\right) \qquad \textbf{(1.13)}$$

1.3.4.4. Cluster Analysis

Cluster analysis is a technique employed to find the most populated configurations during a simulation.[55] The process of obtaining clusters from a molecular dynamics trajectory involves first calculating the root mean squared deviation (RMSD) of C_α atoms between all pairs of aligned structures.[56] The RMSD is a measure of the average distance between the atoms of superimposed molecules, as seen in Equation 1.14.[50] In this equation, N is the total number of atoms and δ_i is the distance between corresponding C_α atoms of superimposed structures. The cluster analysis process involves calculating for each structure the total number of other structures, or "neighbour conformations," that fall within the specified RMSD cut-off. The structure with the highest number of neighbour conformations is taken as the center of the cluster. This center configuration is then combined with all of its neighbour conformations to form the first cluster. The structures of this cluster are then eliminated from the pool of structures. This process is repeated on the smaller pool of structures and continues until the specified number of clusters is

met. The assumption is that the clusters that are more populated are more likely to be the native structure.

$$RMSD = \sqrt{\sum_{1}^{N} \frac{1}{N} (\delta_i)^2} \quad \textbf{(1.14)}$$

1.4. Methods to Determine Protein Structure

1.4.1. Introduction

Several popular methods that are employed to experimentally investigate protein structure include circular dichroism (CD),[57–59] electron microscopy,[60–62] X-ray crystallography,[63,64] and nuclear magnetic resonance (NMR).[65–67] CD and EM are both low resolution methods, where CD can provide overall secondary structure content of proteins and EM can give a low resolution image of the overall shape of large proteins. On the other hand, both X-ray crystallography and NMR can provide atomic resolution data of the three dimension structure of proteins.

The structure of a protein is determined by consolidating experimental data with its known amino acid sequence and knowledge of the preferred geometry of atoms in amino acids. The experimental data required includes the diffraction pattern for X-ray crystallography; the difference in absorbance between left and right circular polarized light for CD; the distance between pairs of atoms that are close in space for NMR; and the image of the overall shape for electron microscopy. The method of choice largely depends on the protein studied (size, shape, nature, and difficulty of sample preparation) and the desired output information (structure, dynamics, shape, or binding). In this thesis, the structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ has thus far only been investigated using circular dichroism. However, some theory concerning EM, X-ray crystallography, and NMR are discussed below because these are common methods in the field of protein structural studies. In order to experimentally investigate the structure of a protein, researchers first require a means of producing the protein of interest.[68–70]

1.4.2. Protein Expression

Researchers have developed numerous techniques in order to manufacture proteins in the laboratory. One method developed to produce short proteins and peptides is chemical synthesis, which can make small amounts of very pure peptides.[71] However, chemical synthesis is not a viable option for large, complex proteins because the reaction is limited by the synthetic efficiency of each reaction step, and unfortunately each amino-acid must be added in a stepwise fashion.[71] For example, a 100 residue peptide with a 97% efficiency for each step provides an overall 5% yield.[71] Another popular technique is recombinant protein expression, where living cells can be harnessed as factories to produce proteins of interest based on a supplied DNA template.[68] Proteins that are produced from DNA templates are referred to as recombinant protein.

Recombinant protein expression is accomplished using a host system and DNA template. The host can be bacteria, yeast, insect, or mammalian cells, while the DNA source can be a virus, plasmid, or artificial chromosome.[68] This thesis focuses on employing the most widely used and developed expression system which involves using *Escherichia coli* (*E. coli*) bacterial host cells with a plasmid DNA template.[69] The *E. coli* - plasmid expression system is very popular because bacteria are easy to culture, grow rapidly, and can produce high yields of recombinant protein, while the plasmid is easy to manipulate genetically.[69]

Figure 1.6 shows the process of recombinant protein expression using an *E. coli* - plasmid expression system. A plasmid is a circular piece of DNA that contains the genetic template (construct) that codes for the protein of interest.[70] Through heat shock, the cell membrane becomes more permeable, allowing the plasmid to be transformed into the *E. coli*. The cells are cultured and induced to overexpress the desired protein.[70] During expression, the protein can either remain soluble, associate with the cell membrane, or become insoluble as inclusion bodies. The cells are then lysed to disrupt its membrane wall, and the protein is released into a medium that prevents degradation of the protein. After locating and identifying the protein, usually through gel electrophoresis and western blotting, at least two purification steps are generally required to separate the protein of interest from other cellular and protein impurities. Once

sufficient quantities of relatively pure protein has been acquired, the protein sample can be prepared for structural studies.

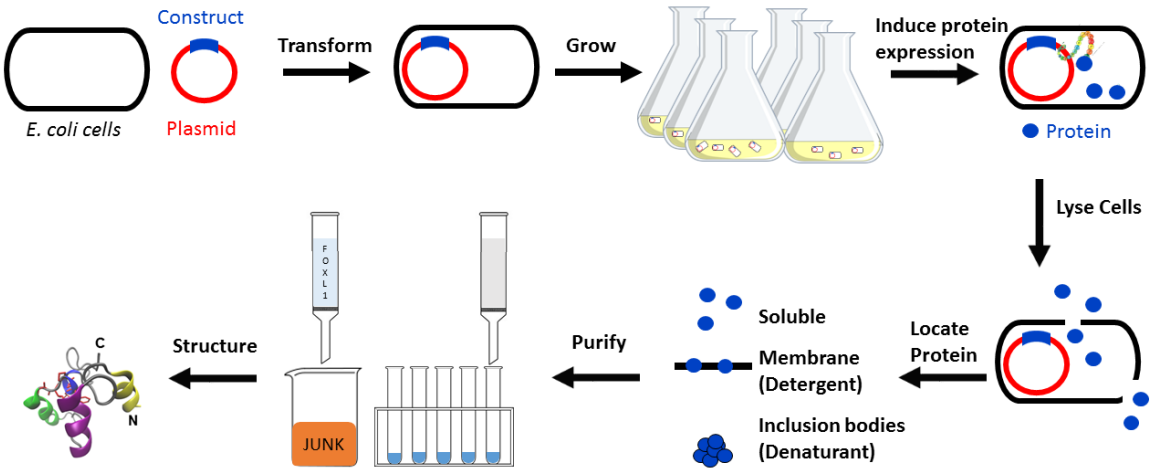


Figure 1.6: Recombinant protein expression using an *E. coli* - plasmid expression system. This process involves transforming a plasmid into *E. coli* cells, growing the *E. coli* cells, and then inducing protein expression. Lysing the cells disrupts the membrane and the protein can be recovered. The protein can remain soluble, associate with the membrane, or form insoluble inclusion bodies. After purification, structural investigation of the protein can begin.

One problem often faced during recombinant protein expression is that there is little or no expression of the foreign gene.[69] In order to increase the quantity of protein expressed, numerous variables can be changed or manipulated such as the strain of the *E. coli* and the DNA template.[68] First of all, there are many different types of genetically altered *E. coli* strains that were developed to express foreign genes. One *E. coli* strain may be more suited than another to express a particular foreign gene.[69] Table 1.1 list a few strains of *E. coli* and the purpose for which they were developed. These were the strains investigated to express $FOXL1_{CTERM/MUT}$. A second variable that can be manipulated is the DNA template. If a protein gene is difficult to express individually, then the DNA template can be modified to contain the protein gene fused to one or more other protein/peptide genes that stabilize the expression.[68,69] These so-called "tags" can increase yields, facilitate protein detection/identification, increase solubility, and assist with purification of the protein.[68,69] Table 1.2 lists a few tags and their purpose. These tags were employed to assist with the expression, identification, and purification of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$.

Table 1.1: Several strains of *E. coli* used in recombinant protein expression.[69]

***E. coli* Strain**	**Purpose**
C41(DE3)	Overexpressing toxic protein
C43(DE3)	Express a gene which codes for membrane and globular protein
BL21(DE3) pLys S	Unable to degrade foreign recombinant protein because of two deleted protease genes (Lon and OmpT) Controlled expression to prevent leaky expression of toxic gene

Table 1.2: Some common tags used in recombinant protein expression.[68,69]

Tag	**Purpose**
6His-tag	Affinity tag composed of six consecutive histidine residues used for purifying recombinant proteins and detection via western blot
S-tag	Affinity tag composed of a 15-residue peptide used for purification and colorimetric detection via western blot
SN-fusion	Staphylococcus aureus nuclease (SN) fusion protein employed to stabilize expression of the gene

One of the most important techniques for isolating and purifying proteins is chromatography.[70] Two commonly employed chromatography techniques are size exclusion gel filtration and affinity chromatography.[70] In size exclusion gel filtration, a mixture of molecules are separated on the basis of size. In this technique, a sample is passed through a column of porous beads. Since larger analytes cannot enter as many porous beads, they experience a smaller volume path to travel and are eluted first. In affinity chromatography, an analyte that has an affinity for a column binds to it while the other impurities are eluted from the column. The protein of interest is then eluted from the column by addition of a reagent that competes for the binding sites. For example, a Ni affinity columns can bind a 6His-tag containing protein for purification.[69,70] The negatively charged His residues interact with Ni^{2+} that has been immobilized on a matrix. The 6His-tagged protein remains on the column, while low concentrations of imidazole is employed to wash the column to remove impurities. The protein can be recovered by eluting the column with an imidazole solution with an increasing concentration gradient.

Two techniques that are employed to identify and characterize the composition of a protein mixture are gel electrophoresis[72] and western blotting[73] as shown in Figure 1.7. Gel electrophoresis is a technique that yields information about the size, purity, and relative amount of proteins in a sample.[70] To perform gel electrophoresis, samples are loaded into lanes on a gel,

where one of the lanes is reserved for a reference ladder. Under the influence of an electric current, the proteins migrate down the gel, with lower molecular weight proteins migrating faster than larger proteins. Upon staining the gel (i.e. Coomassie or silver staining), the protein bands become visible. The size of the protein can be assessed by comparing a band to the reference ladder, which has colored bands of known molecular weight. If a lane has more than a single band, then that sample is impure. Also, larger bands on a gel have more protein than smaller bands, which gives an indication of the relative amounts of protein in a sample. However, if the concentration of a protein sample is too low, it may not be visible on the gel.

Although electrophoresis is an effective technique to determine the size, purity, and relative amount of protein in a sample, it provides no experimental data to prove the identity of any of the protein bands besides their approximate molecular weight. On the other hand, western blotting is a technique that enables a protein to be positively identified. In this method, gel electrophoresis is run and the sample is then transferred from the gel to a synthetic membrane.[73] The transferred blots can then be probed with antibodies that bind only to specific proteins in order to confirm the identity of the protein of interest. To enable easy identification via western blot, recombinant proteins are often designed to have a peptide tag, such as the His-tag, that binds to commercially available antibodies.[73]

Once sufficient quantities of relatively pure protein has been acquired, the protein sample can be prepared for structural investigation by circular dichroism, electron diffraction, X-ray crystallography or nuclear magnetic resonance.

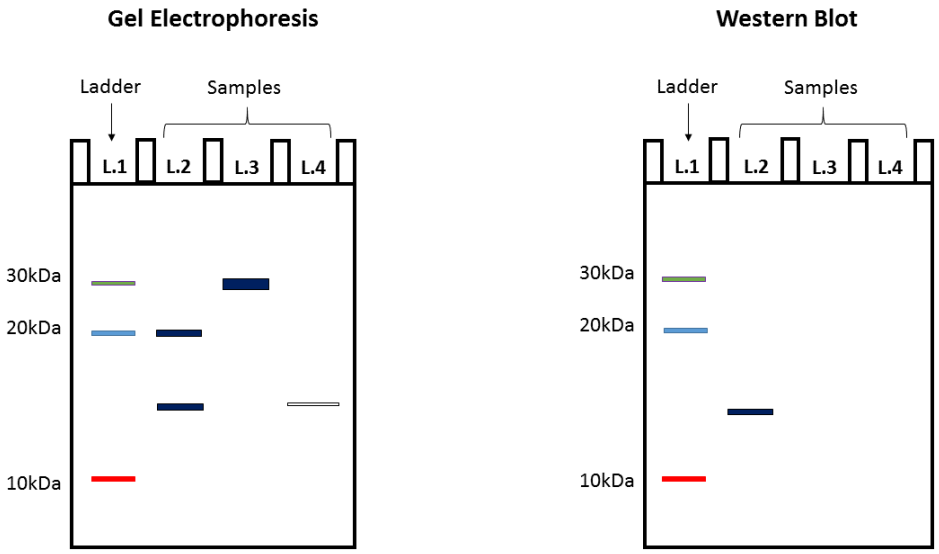


Figure 1.7: Two techniques used to characterize protein samples are (left) gel electrophoresis and (right) western blot. Gel electrophoresis allows the size, purity, and relative quantities of proteins in each sample to be determined. For example, lane 1 (L.1) shows a colored referenced ladder where the bands have molecular weight of 10 kDa (red), 20 kDa (blue), and 30 kDa (grey); lane 2 (L.2) is an impure sample because it has two bands of similar amounts at approximately 15 kDa and 20 kDa; lane 3 (L.3) is a potentially pure sample with a protein of 30 kDa; lane 4 (L.4) reveals that if a protein sample is not concentrated enough, it will not show up on a gel upon staining. A western blot probes for the protein of interest using antibodies, and only proteins that interact with the antibodies appear.

1.4.3. Circular Dichroism

Circular dichroism (CD) is a spectroscopic technique that can be employed to analyze the secondary structure content of a protein.[57–59] This technique is also used to determine how factors like mutations, temperature, pH, ionic strength, ligands, denaturants, and binding interactions affect the conformation and stability of a protein.[57,58] CD is particularly advantageous in protein studies because it is a rapid, non-destructive technique that only requires a small amount of proteins (~ 20 µg) in physiological buffers for secondary structural characterization.

The circular dichroism phenomenon in proteins originates from the carbon atom that is adjacent to the peptide bond, referred to as C_α.[57,59] The C_α atom is bonded to four different constituents in all amino acids (except glycine), making it a chiral center and asymmetric.[59] The chirality of the C_α atom creates asymmetry in the peptide bond chromophore.[59] When left and right circular polarized light passes through a protein sample, the amide bond undergoes an

electronic transition, either $n \rightarrow \pi^*$ or $\pi_o \rightarrow \pi^*$, that absorbs light between 215 – 230 nm or 185 – 200 nm, respectively.[57] These individual peptide bond chromophores can couple differently with each other based on the secondary structure conformation of the protein, inducing an overall chirality which causes different amounts of left of right circularly polarized light to be absorbed.[59] CD measures the difference between the absorption of left and right circularly polarized light over a range of wavelengths to yield a spectrum, the shape of which is sensitive to the secondary structure of a protein.

Figure 1.8 shows a representative shape of a CD curve based on the secondary structure of a protein. Proteins that are primarily α-helical show characteristic negative bands at 222 nm and 208 nm, and a positive band at 193 nm. Anti-parallel β-sheet proteins have a negative band at 218 nm and a positive band at 195 nm. Randomly coiled proteins (such as denatured or intrinsically disordered proteins) have a negative band at 195 nm and show low ellipticity at wavelengths greater than 210 nm. In the case of a protein with multiple secondary structure motifs, we assume to a first approximation that the resultant spectra is a linear combination of the individual structures.[57,58]

Several sample requirements for CD concern its purity, solvent, and concentration. To obtain a reliable CD spectrum, the sample should be 95% pure as measured by gel electrophoresis.[58] Secondly, it is imperative to avoid solvents that strongly absorb light in the acquisition range (190 nm – 260 nm).[58] Thirdly, solids suspended in the solvent can scatter the light and therefore must be removed. Finally, the concentration of the protein should be known accurately in order to both acquire a representative CD spectra and estimate the secondary structure.

Although CD provides secondary structure information, it does not yield residue-specific details required for tertiary structure characterization. Tertiary structure information can be obtained from techniques like electron microscopy, X-ray crystallography, or nuclear magnetic resonance.

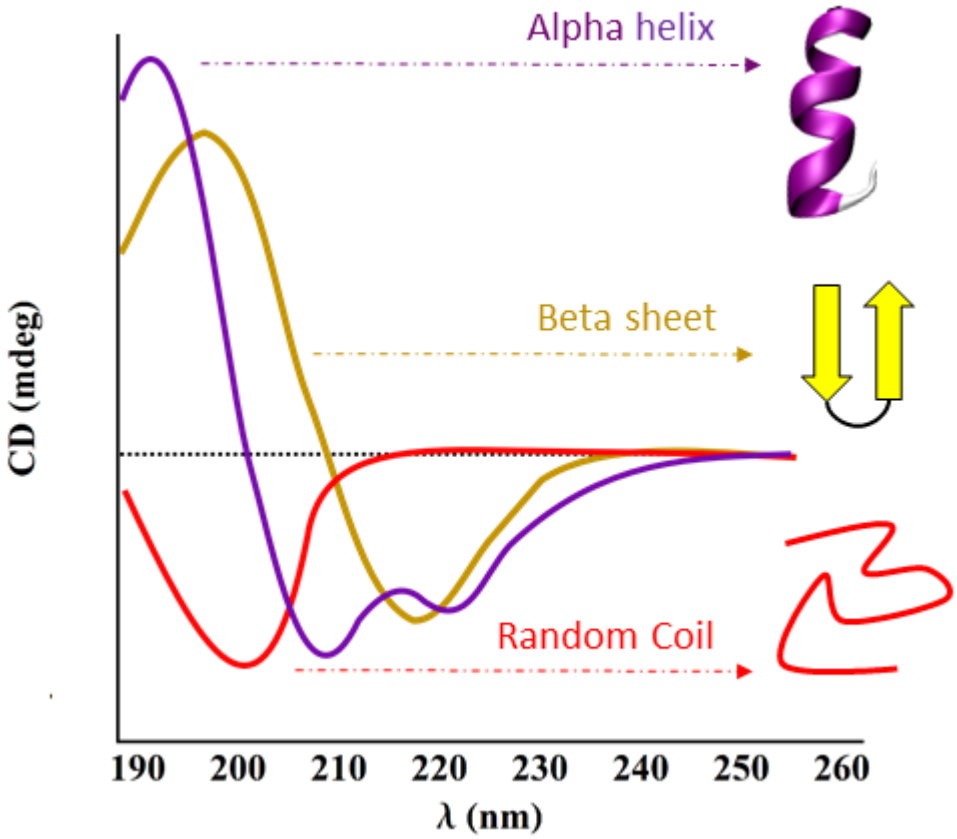


Figure 1.8: Representative circular dichroism spectra showing the differential absorption between left and right circularly polarized light as a function of wavelength for several protein samples. The shape of the curve reveals the secondary structure content of proteins: α-helix has negative minima at 222 and 208 nm, and a positive maximum at 193 nm; β-sheet has a negative minimum at 218 nm and a positive maximum at 195 nm; random coil has a negative minimum at 195 nm and low absorbance greater than 210 nm.

1.4.4. Electron Microscopy

Electron microscopy (EM) is a technique that can be employed to determine the structure of large macromolecular complexes.[60–62] In an EM experiment, a beam of electrons interacts with a specimen to yield 2D images of the biomolecule in a range of orientations.[60] These 2D images can be aligned, averaged, and combined computationally to create a 3D reconstruction of the macromolecule.[61,62]

EM is particularly well suited for structural studies of macromolecules that are difficult to crystalize such as proteins with intrinsically disordered regions.[60] Disordered or flexible regions

are revealed by a reconstructed image that appears fuzzy and unresolved even with increasing number of averaged images. One major advantage of EM over X-ray and NMR is that the output is images as opposed to complex data that must be interpreted. Unfortunately, EM is a destructive method since the electron beam causes radiation damage to biological samples.[60–62] Another downside is that biological samples display low contrast images because the sample is composed of mostly light elements.[60] Also, although tertiary structure information can be extracted from electron microscopy, these experiments rarely produce atomic level detail. EM studies are often combined with experiments that can obtain atomic level detail like X-ray crystallography or nuclear magnetic resonance. One final limitation is that biomolecules must be larger than 100 kDa.[60] Thus, EM is not a viable technique for structural studies for FOXL1 which only has a molecular weight of 36.49 kDa.

1.4.5. X-ray Crystallography

X-ray crystallography is an experimental method that can provide atomic resolution detail of the structure of a protein.[63,64] In this technique, an intense beam of X-rays strike a crystal and are scattered by the electrons of atoms to produce an overall diffraction pattern. The diffraction pattern contains information about the distribution of electrons, which is interpreted to determine the arrangement of atoms in the crystal.[74]

Most of the structures that are listed in the Protein Data Bank (PDB) were determined using X-ray crystallography because it can provide the structures of proteins of any molecular size with a resolution of ~1 - 3Å.[64] This technique requires a crystal that is large (> 0.1 mm diameter) and regular in structure, which does not have significant imperfections like cracks.[63,64,74] X-ray crystallography is an excellent choice for rigid proteins that form ordered crystals, but unfortunately does not work well for flexible, disordered proteins that do not necessarily pack into an ordered, crystalline structure. In fact, disordered regions of crystallized proteins are often hard to observe in the electron density maps determined from X-ray crystallography.[63]

1.4.6. Nuclear Magnetic Resonance

In order to obtain accurate information about protein function, it is important to study their structure as close to the natural state as possible. Nuclear magnetic resonance (NMR) has been

employed to solve the three-dimensional structure of many proteins in their natural solution state.[67,75–78] The studies of proteins by NMR are based on acquiring information concerning through-bond and through-space connectives.[79] NMR techniques based on scalar coupling, the coupling between two nuclei that are connected via bonds, can be used to identify connectivity between nuclei, facilitate resonance assignment, and give information on torsion angles to yield more accurate structures.[80] NMR techniques based on dipolar coupling, the coupling between two nuclei that are close in space, can be used to attain three-dimensional structural information for proteins.[81]

For small proteins and peptides, through-bond and through-space interactions are most conveniently obtained by two-dimensional (2D) NMR methods. Common homonuclear ^{1}H 2D techniques used to acquire scalar coupling information include COSY and TOCSY,[82] and to obtain dipolar coupling correlations are 2D NOESY or 2D ROESY. 2D ^{1}H NMR studies are generally limited to proteins less than 10 kDa.[80] In larger proteins, an overwhelming number of interactions lead to overlap and degeneracy issues in the resulting spectrum, which prevents the unambiguous assignment of the peaks.[77] Consequently, three-dimensional NMR is becoming the method of choice for structural studies of proteins by NMR because of its ability to reduce the extent of spectral overlap by extending 2D peaks into a third dimension. However, for the structural study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$, both of which have a molecular weight less than 10 kDa, the first NMR methods that should be employed to obtain through-bond and through-space connectivity information are 2D TOCSY[82] and 2D NOESY,[81] respectively.

1.4.6.1. TOCSY

Two-dimensional total correlation spectroscopy (2D TOCSY) is a homonuclear ^{1}H NMR technique that uses consecutive scalar couplings to correlate all hydrogens within a spin system, as seen in Figure 1.9. The technique is also known by the acronym HOHAHA for homonuclear-Hartmann-Hahn.[82]

TOCSY is especially applicable to molecules that are intrinsically composed of separated groups of coupled spins.[80] Two nuclei are coupled if they are within three bonds of each other. A spin system is a collection of nuclei (e.g. protons) that are directly or indirectly coupled to each other by consecutive spin coupling. In the case of proteins, each amino acid constitutes one or more

independent spin systems, owing to the carbonyl carbon which breaks any scalar coupling between adjacent amino acids as shown in Figure 1.10. Thus, TOCSY is very applicable to proteins, because each amino acid gives characteristic TOCSY signal patterns, which assists in the identification of resonances on an NMR spectrum.[83]

TOCSY has a few limitations. First of all, TOCSY only shows local connectives within a spin system. As a result, a particular amino acid can be identified on a TOCSY spectrum, but the location of that residue in relation to its amino acid sequence cannot usually be elucidated exclusively from a TOCSY analysis, especially if there are several of the same amino acid in the sequence. Thus, it is usually necessary to supplement TOCSY data with spatial connectivity information.

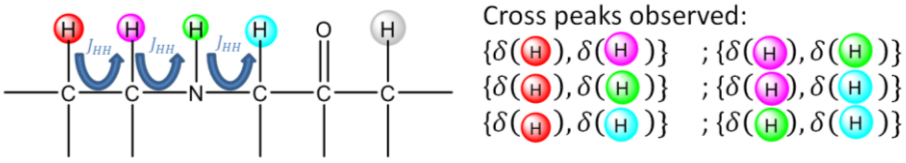


Figure 1.9: Scalar coupling between a network of coupled spins. A TOCSY cross peak is observed between each pair of hydrogens within a spin system. Since the hydrogen in grey is not part of the larger spin system, no cross peaks between it and the other hydrogen are observed.

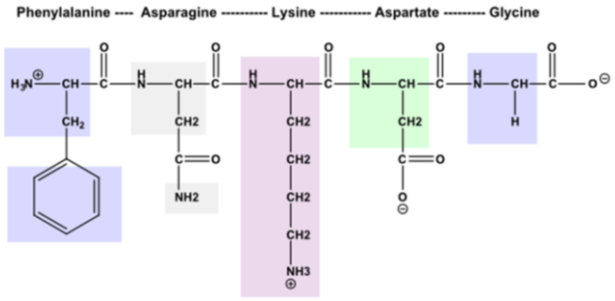


Figure 1.10: Each amino acid in a protein is its own independent spin system. The spin systems for a five residue peptide are highlighted. Some amino acids such as phenylalanine and asparagine contain two spin systems while others like lysine, aspartate, and glycine have a single spin system.

1.4.6.2. NOESY

Two-dimensional nuclear Overhauser effect spectroscopy (2D NOESY) is a homonuclear ^{1}H NMR experiment that correlates protons that are close in space.[81] It is a widely used technique in structural studies of proteins because short internuclear distances can be used to determine how the amino acids are oriented with respect to each other in space.

A 2D NOESY spectrum contains diagonal peaks and symmetrically placed cross peaks.[81] The diagonal peaks result from magnetization that remained on the first spin during the mixing sequence, and thus give no structural information. On the other hand, a cross-peak connecting two signals implies the protons are close in space, usually less than 5 Å.[65] However, in severe spin diffusion cases, where magnetization is exchanged spontaneously between spins, cross-peaks for protons that are greater than 5 Å apart can appear on a NOESY spectrum.[80] As a result, for proteins, the mixing time is kept to a minimum between 100-200 ms to prevent spin diffusion.[80] Other undesirable peaks that may arise are chemical exchange peaks because, like NOEs, they involve a transfer of magnetization between nuclei. As a general note, unlike scalar coupling, dipolar coupling in the solution state does not cause splitting of the signals.[79]

The knowledge of through-space interactions from NOESY data can assist in both the assignment of resonances and the structural determination of proteins.[2,84] NOESY is extremely useful for connecting adjacent amino acids because hydrogen in adjacent spin systems will almost always be close in space.[79] This will help assign the resonance peaks of each residue (obtained from 2D TOCSY) relative to its position in the amino acid sequence. Furthermore, since secondary structural elements show characteristic through-space interactions, a NOESY spectrum can be analysed for the presence or absence of these peaks to determine whether sections of the protein possess these secondary structure elements.[84] In addition, NOEs occurring between hydrogens far apart in the amino acid sequence yield information on how the protein is folded in 3D. Finally, the intensity of a NOESY cross peak is proportional to $1/{r_{AB}}^{6}$, where r_{AB} is the inter-proton distance.[66] Thus, more intense cross peaks generally denote protons that are closer in space.

1.5. Goals

The long term goal of this research is to determine how the deletion of GIPFL in the mutant FOXL1 protein leads to a human genetic disease. Before this objective can be achieved, a structural study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ must first be performed. This thesis sets out to determine the structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ using bioinformatics, computational simulations, and experimental methods. In this thesis, several bioinformatics tools were first employed to predict protein disorder, secondary structure, and sequence alignment of these proteins. As will be shown, the bioinformatics tools predicted that the GIPFL mutation occurs in an ordered, structured, evolutionarily-conserved portion of the C-terminal domain of FOXL1. In addition, computational simulations were also performed that combined the PACE model and replica exchange molecular dynamics to determine the most statistically probable structures of both $FOXL1_{CTERM}$ and $FOXL1_{MUT}$. These computational results predict that the deletion of GIPFL in $FOXL1_{MUT}$ protein disrupts its hydrophobic core and causes it to become more disordered than $FOXL1_{CTERM}$. Finally, experimental progress was made towards expressing and purifying both $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ through recombinant protein expression. Preliminary structural data was obtained using circular dichroism which indicated that $FOXL1_{CTERM}$ has a predominantly helical secondary structure while $FOXL1_{MUT}$ was partially helical with some randomly coiled regions. The combination of bioinformatics, computer simulations, and preliminary experimental results strongly suggests that the GIPFL deletion increases the disorder of the $FOXL1_{MUT}$ protein in comparison to $FOXL1_{CTERM}$.

Chapter 2

Methodology

2.1. Computational

The computer simulations were performed with the VMD software package (version 1.9.1).[85] The structure of the most C-terminal 69 residues of wildtype FOXL1 ($FOXL1_{CTERM}$) and corresponding residues of the deletion mutant ($FOXL1_{MUT}$) were deduced by molecular dynamics simulations. The initial structures of these proteins were generated in an extended state and were coarse-grained (CG) based on the united-atom (UA) model and then solvated in coarse grain water (CGW) molecules with an additional 10 Å of water padding. All non-bonded interactions were shifted to zero between distances of 9 Å and 12 Å. The time-step for all simulations was set to 5 fs. Pair lists were updated at least once per 10 steps, with a 12 Å pair list cut-off. A Langevin thermostat with a damping coefficient of 10 ps^{-1} was employed to maintain temperature. The solvated system was energy minimized with the PACE coarse grain force field.[24,25] The minimized system was then equilibrated for 3.5 ns and then submitted for a 20 ns simulation using the canonical NPT ensemble at 300 K and 1 bar pressure. The constant pressure of 1 atm was

maintained with a Nosé-Hoover Langevin piston barostat[86,87] using a piston period of 100 fs and a decay time of 50 fs. The waters were then removed, and the protein was re-solvated in a cubic box with a side length of 90 Å and neutralized by the addition of a single sodium cation. This system was then minimized, equilibrated (3.5 ns), and submitted for a 5 ns simulation, all using a canonical ensemble at 300 K. In this equilibration, the backbone atoms (C_α, the amide N, and the carbonyl carbon) of the protein were fixed.

Replica exchange molecular dynamics was then employed to fold the protein and sample various configurations. The REMD simulations contained 32 replicas with temperatures ranging from 300 K to 500 K.[††] Each replica started with the same conformation resulting from the molecular dynamics equilibration. Exchanges were attempted every 0.05 ns, having an acceptance rate of about 15%. A 4000 ns replica exchange simulation was performed, where the configurations are saved at every 0.05 ns. The full simulation at 300 K was used for analysis. The dominant structure in the 300 K replica was determined by clustering analysis based on the root mean squared distance (RMSD) of all C_α atoms. Clustering analysis was performed in the VMD 1.9.1 program using a 3 Å RMSD criteria.[55]

2.2. Experimental

2.2.1. General

Three different constructs (1) 6His—$FOXL1_{CTERM/MUT}$, (2) S-tag—6HIS—$FOXL1_{CTERM}$, and (3) SN—$FOXL1_{CTERM/MUT}$—6HIS were investigated in order to determine the structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$. The first construct investigated, 6His-$FOXL1_{CTERM/MUT}$, had low protein expression and significant sample impurity. The second construct studied, S-tag—6HIS—$FOXL1_{CTERM}$, also had low protein expression and sample impurities, but allowed for identification of the target protein by western blot and enabled structural characterization of an impure target protein by circular dichroism. The final construct, SN— $FOXL1_{CTERM/MUT}$—6HIS, had high expression yields, was easily

[††]The temperatures for each of the 32 replicas in kelvin are: 300, 304.98, 310.05, 315.20, 320.44, 325.76, 331.18, 336.68, 342.27, 347.96, 353.74, 359.62, 365.59, 371.67, 377.84, 384.12, 390.50, 396.99, 403.59, 410.29, 417.11, 424.04, 431.08, 438.25, 445.53, 452.93, 460.46, 468.11, 475.88, 483.79, 491.83, 500

identified via western blot, was successfully enriched, and allowed for some structural characterization of the target protein by circular dichroism.

The expression and characterization for the most promising construct (3) SN—FOXL1$_{CTERM/MUT}$—6His is discussed in detail first. Then, the similar (but concise) expression details for construct (1) 6His—FOXL1$_{CTERM/MUT}$ and (2) S-tag—6HIS—FOXL1 $_{CTERM}$, are then discussed.

2.2.2. Construct 3: SN—FOXL1 $_{CTERM/MUT}$—6His

2.2.2.1. Transformation

A gene that coded for the SN—FOXL1$_{CTERM}$—6His (or SN—FOXL1$_{MUT}$—6His) protein construct was synthesized and inserted into the pet29a(+) vector by GenScript (New Jersey, USA). C43(DE) *Escherichia coli* (*E. coli*) overexpress competent cells were removed from the −80 °C freezer and thawed for 20 minutes on wet ice. 1 µL of vector was added to 50 µL of cells on ice and stirred gently with a pipette tip. This sample was incubated on ice for 30 minutes. The cells were then heat shocked in a 42 °C water bath for 45 seconds, and placed back on ice for 2 minutes for recovery. 950 µL of room temperature expression recovery medium - composed of 2xYT (16 g/L tryptone, 10 g/L yeast, and 5 g/L NaCl) - was added to the cells in the culture tube. Tubes were placed in a shaking incubator at 175 rpm for 1 hour at 37 °C. 50 µL and 100 µL samples of transformed cells were spread onto separate agarose plates, which were composed of 2xYT and 30 µg/mL of kanamycin antibiotic. The plasmid gave kanamycin resistance to the cells that underwent transformation, while cells that did not take up the plasmid died. The plates were incubated overnight at 37 °C. One isolated colony was obtained from the agar plate using a sterile toothpick and placed in 5 mL of 2xYT with 30 µg/mL of kanamycin, and then incubated for 5 hours at 37 °C and 175 rpm.

2.2.2.2. Stock Sample Preparation

Stock samples of successfully transformed bacteria were made by combining 500 µL of bacteria cells with 500 µL of glycerol into Eppendorf tubes. These samples were stored at −80 °C. Subsequent protein expression started from these stocks.

2.2.2.3. Protein Expression

An overnight culture was prepared containing 500 µL of the transformed sample (or one of the stock samples from Section 2.2.2.2), 30 µg/mL of kanamycin, and 75 mL of sterile 2xYT, and then incubated overnight at 175 rpm and 37 °C. 10 mL of overnight culture was added to each of 6×4 L conical flasks containing 1 L of sterile 2xYT with 30 µg/mL of kanamycin. The sample was incubated at 175 rpm and 37 °C until a UV absorbance of approximately 0.6 at 600nm was reached (~2.5 hours). Then, 500 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to each 1 L flask to induce cells to make protein. The sample was incubated for 3 hours at 175 rpm and 37 °C. Mixture was spun in centrifuge at 5000 rpm for 30 minutes. The media was discarded and the pellet was collected. 25mL of 5 mM phenylmethylsulfonyl fluoride (PMSF - a serine protease inhibitor) in tris-buffered saline (TBS) was added to the pellet.

2.2.2.4. Cell Lysis

A French press or sonication was employed to disrupt the bacterial cell membrane and recover the protein. For cell lysis via French press, the bacteria cells were ran through a French press three times using a pressure of 12000 psi to disrupt the cells. The cells were collected in a chilled Erlenmeyer flask. In order to disrupt cells via sonication, the sample was sonicated fifteen times for 20 seconds, with one minute breaks between each cycle. Cells were kept on ice for the entire duration.

2.2.2.5. Locate Protein

A successfully expressed protein can be (a) in the supernatant (b) associated with the membrane, or (c) in inclusion bodies. In order to determine the protein's localization, the lysate was separated into these three fractions and the protein visualized. In this case, the expressed protein was found in inclusion bodies.

(a) **Supernatant:** The cell lysate mixture was centrifuged for 30 minutes at 4 °C and the soluble layer was decanted off the pellet. The remaining pellet was kept for the "associated with membrane" sample below. To the supernatant, 5 g of DE52 (diethylaminoethyl, anion exchange pre-swollen whatman cellulose resin) that was equilibrated in TBS was added and spun for 1 hour at 4 °C. Sample was centrifuged

for 20 minutes at 4 °C and 10000 rpm. The supernatant was decanted off the DE52 and the resin was discarded. This solution is the “crude soluble” layer discussed later in the results section.

(b) **Associated with the membrane:** To the pellet from (a), 25 mL of 0.2% (w/v) 3-(3-(cholampidopropyl) dimethylammonio)-1-propanesulfonate (CHAPS) detergent in TBS was added and spun in the cold room for 1 hour. Sample was centrifuged for 20 minutes at 4 °C and 10000 rpm. Detergent soluble layer was decanted off. The remaining pellet was kept for the “inclusion bodies” below. To the solution, 5 g of DE52 equilibrated in detergent solution was added and spun for 1 hour at 4 °C. The detergent soluble layer was decanted off DE52 and the resin was discarded. At this point, this solution is the “crude, associated with the membrane” sample discussed later in the results section.

(c) **Inclusion bodies:** To the pellet from (b), 25 mL of 6 M urea, 0.2% CHAPS, in TBS was added and spun in the cold room for 1 hour. Sample was centrifuged for 20 minutes at 4 °C and 10000 rpm. To this solution, 5 g of DE52 equilibrated in 6 M urea, 0.2% CHAPS, in TBS solution was added and then spun for 1 hour at room temperature. The soluble layer of the centrifuged material was decanted off DE52 and the resin was discarded. At this point, this solution is the “crude inclusion body” sample discussed later in the results section.

2.2.2.6. Ni Affinity Column Purification

To initially purify the crude samples, nickel-bound immobilized metal ion affinity chromatography (IMAC) was employed. This is done in order to separate the expressed protein (SN—FOXL1$_{CTERM}$—6His) from other cellular and protein impurities.

First, a Ni column was prepared and equilibrated with one of the buffers listed in Table 2.1. Imidazole (imdzl) was added to the crude sample from Section 2.2.2.5 to match the imidazole concentration in the equilibration buffer. The sample was then run through the prepared Ni column and the flow-through was collected. The target protein binds to the Ni column via the attached 6His-tag, while many other impurities elute in the flow-through and wash. 25 mL of

starting buffer was used to wash the column, which was collected as 3×8.3 mL washes. The target protein was then eluted from the Ni column with increasing concentrations of imidazole using one of the elution schemes (detailed in Table 2.1) and collected as ~1 mL fractions.

Table 2.1: Several elution schemes used for Ni affinity column purification

No.	Ni column details
1	**Initial column equilibration**: 5 mM imidazole in 6 M urea and TBS **Elution**: 10 mL of each of 5, 10, 20, 50, 100, 200, and 300 mM imidazole in 6 M urea and TBS
2	**Initial column equilibration**: 5 mM imidazole in 6 M urea, 0.2% CHAPS, TBS **Elution**: 10 mL of each of 5, 10, 20, 50, 100, 200, and 300 mM imidazole in 6 M urea, 2% CHAPS, TBS
3	**Initial column equilibration**: 5 mM imidazole in TBS **Elution**: 10 mL of each of 5, 10, 20, 50, 100, 200, and 300 mM imidazole in TBS
4	**Initial column equilibration:** 15 mM imidazole in 6 M urea, 0.2% CHAPS, TBS **Elution**: 10 mL of each of 15, 25, 50, 75, 100, 200, and 300 mM imidazole in 0.2% CHAPS, 6 M urea, TBS
5	**Initial column equilibration:** 20 mM imidazole in 6 M urea, 0.2% CHAPS, TBS **Elution**: 10 mL of each of 20, 50, 100, 200, and 300 mM imidazole in 6 M urea, 0.2% CHAPS, TBS
6	**Initial column equilibration**: 20 mM imidazole in 6 M urea, 0.2% CHAPS, TBS **Elution**: 10 mL × 20 mM; 20 mL × 50 mM; 20 mL × 100 mM, 10 mL × 200 mM, 10 mL × 300 mM imidazole in 6 M urea, 0.2% CHAPS, and TBS
7	**Initial column equilibration**: 20 mM imidazole in 6 M urea and TBS **Elution:** 10 mL of 5, 10, 20, 50, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 280, 290, 200, 300 mM of imidazole in 6 M urea and TBS
8	**Initial column equilibration**: 5 mM imidazole in 0.2% CHAPS and TBS **Elution:** 10 mL of 5, 10, 20, 50, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 280, 290, 200, 300 mM of imidazole in 0.2% CHAPS and TBS

2.2.2.7. UV Detection

The UV-vis absorbance of the eluted fractions was measured at 280 nm using the relevant solvent as a blank. The 280 nm wavelength probes for aromatic amino acids in proteins including tryptophan, tyrosine, and phenylalanine.[70] Peaks in a plot of "absorbance vs. fraction number" generally indicates the presence of protein.

The concentration (in mg/mL) of a protein can be approximated using Equation 2.1, where A_{280} is the UV absorbance at 280 nm, MM is the molar mass of the protein, l is the path length

of the cuvette (usually 1 cm), and ε_{280} is the molar absorptivity ($M^{-1}cm^{-1}$) of the protein at 280 nm. The molar absorptivity (ε_{280}) can be approximated from the number of tryptophan (n_{Trp}), tyrosine (n_{Tyr}), and disulfide bonds (n_{s-s}) in a protein using Equation 2.2.[70]

$$C = \frac{A_{280} \times MM}{\varepsilon_{280} \times l} \quad \textbf{(2.1)}$$

$$\varepsilon_{280} = 5550 \times n_{Trp} + 1490 \times n_{Tyr} + 125 \times n_{s-s} \quad \textbf{(2.2)}$$

2.2.2.8. Gel Electrophoresis

Tris-tricine polyacrylamide gel electrophoresis (PAGE) was employed to determine if the protein was successfully expressed and to locate the protein. The most promising fractions were first determined by UV-vis. 50 µL of each fraction was combined with 25 µL of tricine-sample buffer in an Eppendorf tube and heated for 5 minutes in boiling water. 40 µL of each sample was loaded into the lanes of a 16.5% tris-tricine gel (unless otherwise specified). The gel was run at 90 V until the tricine dye was ~1 cm from the end of the gel. The gel was then Coomassie or silver stained.

a) **Coomassie stained:** The gel was stained using Coomassie-blue for one hour and then de-stained until protein bands appeared using a 200 mL solution composed of 20 mL of ethanol, 14 mL of glacial acetic acid, and 166 mL of distilled water.

b) **Silver stained:** The gel was placed for 20 minutes in fixing solution, which is composed of 50 mL of methanol, 10 mL of glacial acetic acid, 10 mL of fixative enhancer concentrate and 30 mL of distilled water. Then the gel was washed 2×10 minutes with distilled water. The gel was then silver stained until bands appeared using a solution of 17.5 mL of distilled water, 2.5 mL of silver complex solution, 2.5 mL of reduction moderator solution, 2.5 mL of image development reagent, and 25 mL of room temperature development accelerator solution. After staining, 25 mL of 5% glacial acetic acid was used for 15 minutes to stop the reaction.

2.2.2.9. Western Blot

Gel electrophoresis (without staining) of a sample was first run using an electrophorator. A polyvinylidene fluoride (PVDF) membrane was activated by placing it in methanol for 5 seconds, then rinsed twice with distilled water. The gel and PVDF membrane were both equilibrated for 30 minutes in 10 mM of 100 mL of 3-(cyclohexylamino)-1 propane sulfonic acid (CAPS) transfer buffer that had 10% (v/v) methanol. Then, six chromatography sheets (8 cm × 6 cm) were cut and equilibrated in the transfer buffer. The stacking rack was prepared with the gel and membrane in the center, three chromatography papers on both sides, and sponges (soaked with transfer buffer) on the outside. The stacking rack was then placed in the western blot transfer apparatus and filled with transfer buffer. The western blot transfer apparatus was then employed to transfer protein from the gel to the membrane at a low voltage of 50 V for 2 hours. PDVF membrane was placed in blocking reagent (3% skim milk powder in Tween tris buffered saline (TTBS), where TTBS is composed of 50 mM Tris HCl, 150 mM NaCl, and 0.1% Tween 20, at pH 7.4) for over one hour. The blocking reagent was replaced with primary antibody and incubated overnight on a shaker. The membrane was washed with 3×25 mL of TTBS for 5 minutes each with gentle agitation. Secondary antibody was added and then placed on a shaker for 2 hours. The membrane was washed with 3×25 mL of TTBS for 5 minutes each with shaking. Then a color reaction or chemi-luminescence reaction was employed to detect protein blots. The type of primary, secondary, and detection reaction used depends on the type of western blot ran. The different types of western blots discussed in this manuscript are detailed below.

a) **His-tag western:** The primary antibody was a 1:3000 (v/v) dilution of monoclonal anti-polyhistidine antibody found in mouse (Sigma, St. Louis, MO) in 1.25% milk powder in TTBS. The secondary antibody was a 1:5000 (v/v) dilution of anti-mouse IgG (whole molecule) in alkaline phosphatase conjugate developed in goat (Sigma, St. Louis, MO) in 30 mL of TTBS. To perform the color reaction, the membrane was equilibrated for 10 minutes with 20 mL of 0.1 M $NaHCO_3$, 1 mM $MgCl_2$ at pH=9.8. The substrate was prepared in the dark, and contained 60 µL of BCIP (5-bromo-4-chloro-3-indolyl phosphate), 120 µL of NBT (nitro blue tetrazolium chloride) in 0.1 M $NaHCO_3$, 1 mM $MgCl_2$ at pH=9.8. Substrate was added to membrane and covered in tinfoil until bands appeared. TE buffer

at pH 8 (10 mM Tris, 1 mM ethylenediaminetetraacetic acid (EDTA)) was added to membrane to stop the reaction.

b) **SN western:** The primary antibody was a 1:1000 (v/v) dilution of Staphylococcus aureus nuclease (SN) antibody produced in rabbit (MyBioSource, San Diego, CA) in 1.25% milk powder in TTBS. The secondary antibody was a 1:2000 dilution of anti-rabbit IgG produced in goat (Sigma-Aldrich, St. Louis, MO) in 20 mL of TTBS. To perform the color reaction, the membrane was equilibrated for 10 minutes with 20 mL of 0.1 M $NaHCO_3$, 1 mM $MgCl_2$ at pH=9.8. Substrate was prepared in the dark, and contained 60 µL of BCIP, 120 µL of NBT in 0.1 M $NaHCO_3$, 1 mM $MgCl_2$ at pH=9.8. Substrate was added to membrane and covered in tinfoil until bands appeared. NOTE: This happened very quickly within 5-15 seconds. TE at pH 8 was added to membrane to stop the reaction.

c) **S-tag western** (used for construct 2): The primary antibody was a 1:1000 (v/v) dilution of anti-S-tag antibody produced in rabbit mixed with 1.25% milk powder in TTBS. The secondary antibody was a 1:2000 (v/v) dilution of donkey anti-rabbit antibody in TTBS. A substrate containing 500 µL of luminol and 500 µL of hydrogen peroxide were mixed in the dark. The substrate was pipetted on the membrane and left standing in light for 2 minutes. The membrane was imaged using a camera capable of chemi-luminescence detection.

2.2.2.10. Dialysis

By combining information from UV-vis, gel electrophoresis, and western blot, the fractions containing the expressed protein were combined for subsequent dialysis. The combined samples were dialyzed using 1000 kDa molecular weight cut-off (MWCO) tubing against 4 L of distilled water for 12 hours at 4 °C using gentle stirring.

2.2.2.11. Lyophilisation

The solution was flash-frozen using liquid nitrogen and then freeze dried for 12 hours using a Labconco freeze drier (Kansas City, MO, USA).

2.2.2.12. Cyanogen Bromide Digest

In order to prevent the α-helical SN-fusion protein from influencing the structure of $FOXL1_{CTERM/MUT}$, a CNBr digest was employed to cleave SN off the expressed protein (SN—$FOXL1_{CTERM}$—6His or SN—$FOXL1_{MUT}$—6His). The digestion occurred at a methionine residue between the SN and target protein. 1 mL of 70% formic acid was added to the freeze dried sample in small glass vial. One crystal of cyanogen bromide (CNBr) that was ~2 mm in diameter was carefully added to the sample in the fumehood. The glass vial was capped and wrapped in tinfoil for 24 hours (determined by optimization).

One experiment performed involved determining the optimal time for CNBr digestion. In this case, during the CNBr digest described above, 25 µL aliquots were acquired every 8 hours for a total of 3 days, quenched with 500 µL of water, and freeze dried. The solid was dissolved in 60 µL of tricine sample buffer, heated for five minutes, and then 40 µL was loaded onto a gel for gel electrophoresis. The optimal time was determined to be 24 hours.

After CNBr digest, 15 mL of distilled water was then added to quench the reaction. Sample was transferred to a 150 mL beaker for lyophilisation (see Section 2.2.2.11). The solid was then dissolved in 1 mL of 5 mM imidazole in 6 M urea and TBS, and another Ni affinity column was run to separate the SN fusion and target $FOXL1_{CTERM/MUT}$—6His protein (see Section 2.2.2.6). UV-vis (Section 2.2.2.7) and gel electrophoresis (Section 2.2.2.8) was run to determine where the target protein eluted.

2.2.2.13. Size Exclusion Gel Filtration

The running buffer (1 L of 0.2% CHAPS, 6 M urea, and TBS) was degassed under vacuum filtration using a G6 filter paper. A size exclusion column (S-100, HiPrep 16/60) was equilibrated with 2 column volumes of buffer using a flow rate of 0.25 mL/min. A 1.5 mL sample was loaded on the column and fractions were eluted using a flow rate of ~0.2 mL/min (unless otherwise specified). Size exclusion gel filtration was run until 100 × 1.5 mL fractions were collected. UV-vis (Section 2.2.2.7) and gel electrophoresis (Section 2.2.2.8) was run on sample. Fractions containing exclusively the target protein were combined, dialyzed against water, freeze dried, and prepared for structural determination by circular dichroism.

2.2.2.14. Circular Dichroism

Structural studies were done using circular dichroism which was performed on a Jasco J-810 spectropolarimeter (Jasco Inc., Easton, MD) in the far ultraviolet range (190-260 nm) with a 0.5 mm quartz cuvette at room temperature where 20 scans were averaged (unless otherwise specified).

2.2.3. Construct 1: 6His—FOXL1$_{CTERM/MUT}$

A similar transformation protocol as seen in Section 2.2.2.1 was followed to transform the pET-15b plasmid containing FOXL1$_{CTERM}$ (or FOXL1$_{MUT}$) with C-terminal 6His into C43(DE3) *E. coli* cells. The cells were cultured in the presence of 35 µg/mL chloramphenicol and 50 µg/mL ampicillin antibiotics. The cells were induced to express protein following Section 2.2.2.3 and then lysed using the French press procedure seen in Section 2.2.2.4. The cell lysate mixture was centrifuged for 30 minutes at 4 °C and the soluble layer was decanted off the pellet. The desired protein was assumed to be present in the pellet containing inclusion bodies, and thus was recovered using 6 M urea in TBS (both with and without 0.5% N-lauroyl sarcosine detergent). The target protein was never successfully identified or purified due to low expression yields.

2.2.4. Construct 2: S-tag—6His—FOXL1$_{CTERM}$

A similar transformation protocol as seen in Section 2.2.2.1 was followed to transform the pET-29a(+) plasmid containing FOXL1$_{CTERM}$ with N-terminal S- and 6His- tags into BL21(DE)pLysS *E. coli* cells. The cells were cultured in the presence of 35 µg/mL chloramphenicol and 30 µg/mL kanamycin antibiotics. The cells were induced to express protein following Section 2.2.2.3 and then lysed using the French press procedure seen in Section 2.2.2.4. An S-tag western blot (see Section 2.2.2.9) was then run to determine whether the protein was found in (a) the supernatant (b) associated with the membrane, or (c) in inclusion bodies sample. Sample was then purified by nickel affinity chromatography (Section 2.2.2.6). Dialysis (Section 2.2.2.10) was employed to remove unwanted salts and detergents, and samples were concentrated using lyophilisation (Section 2.2.2.11). The sample was then prepared for structural studies using circular dichroism (Section 2.2.2.14).

Chapter 3

Results

3.1. Bioinformatics

In this section, various bioinformatics programs are employed to predict the disorder, sequence alignment, and secondary structure of FOXL1.

3.1.1. Protein Disorder

The ordered and intrinsically disordered regions of FOXL1 was computationally predicted by PONDR-FIT as shown in Figure 3.1. The prediction results of the N-terminal domain are analyzed first. From literature, it is known that a structured N-terminal DNA binding domain spans residue 49-139.[14] Although PONDR-FIT does correctly predict an ordered region from residue 52-128, the N-terminal domain is underestimated on both ends. This is expected since protein disorder

predictors are known to be unreliable along the boundary between the structured and intrinsically disordered regions.[6] We also note the poor predictive accuracy for short disordered regions, as revealed by the two short disordered and one short ordered regions on the N-terminal side of the Forkhead domain. Overall, PONDR-FIT analysis shows good predictive accuracy for the N-terminal domain, except close to the boundary between ordered and disordered regions.

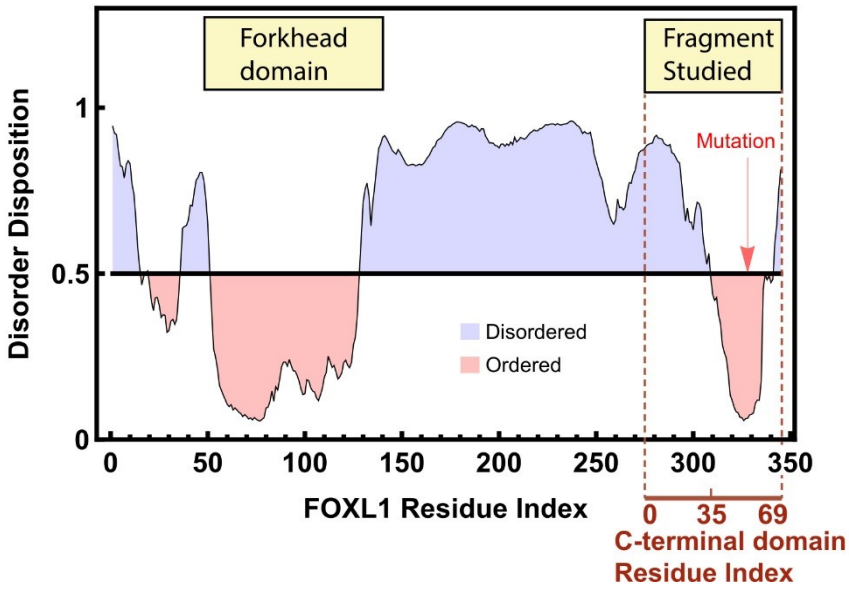


Figure 3.1: Prediction of intrinsically disordered regions in the FOXL1 protein using PONDR-FIT.[6] A disorder disposition of 0.5 represents the threshold between the predicted disordered (> 0.5) and structured (< 0.5) regions. Based on this analysis, a 69 amino acid fragment the spans the ordered C-terminal region was investigated. A mutant FOXL1 protein was also studied, where a five amino acid (GIPFL) segment was deleted. The predicted ordered regions span residue 20-35, 52-128, and 309-345.

PONDR-FIT predicted that the ordered N-terminal domain is connected by a long intrinsically disordered segment (residues 129-308) to an ordered C-terminal region (residues 309-345). This prediction suggests that a C-terminal domain exists for FOXL1 that can maintain its structure and function independently of the rest of the protein. Since it is both computationally and experimentally expensive to study the structure of the full 345 amino acid FOXL1 protein, we want to reduce the size of the system studied. Based on this analysis, a 69 amino acid fragment was studied (residue 277-345) that completely encompasses the predicted ordered C-terminal region and contains a generous portion of the intrinsically disordered domain to account for poor

boundary accuracy. From here on, the residues 277-345 will be denoted as $FOXL1_{CTERM}$ and the same segment with GIPFL deleted (residues 326-330) will be denoted $FOXL1_{MUT}$.

One key insight predicted by PONDR-FIT is that deletion of the GIPFL residues in FOXL1 occurs in the most ordered portion of the C-terminal domain of FOXL1. This suggests that this mutation could severely affect the structure and function of the $FOXL1_{CTERM}$.

3.1.2. Sequence Alignment

A multiple sequence alignment of $FOXL1_{CTERM}$ with twenty-five FoxL1 proteins (from other chordates) was performed by BLASTP, which is illustrated in Figure 3.2. The sequence alignment identified a region of similarity spanning residue 30-69 of $FOXL1_{CTERM}$. In this region, the majority of the aligned amino acids are identical or chemically similar. Therefore, this highly conserved region of $FOXL1_{CTERM}$ is likely to be structurally or functionally important. In particular, the GIPFL residues (50-54) are all identical except phenylalanine ("F") which is still chemically similar to the other amino acids in the alignment. From this information, it is understandable that the deletion of GIPFL arising in $FOXL1_{MUT}$ makes the FOXL1 protein malfunction, leading to a human genetic disease. Structural changes are also possible considering that five highly conserved amino acids are deleted.

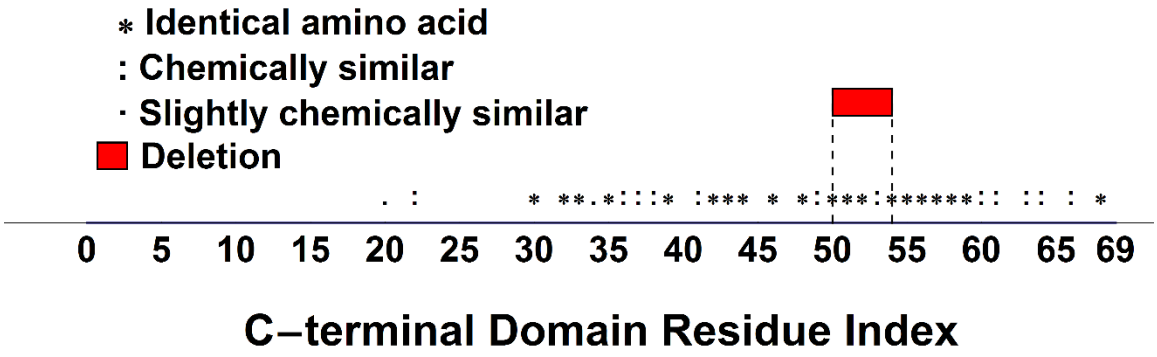


Figure 3.2: BLASTP sequence alignment of $FOXL1_{CTERM}$ with 25 other FoxL1 proteins (reference UniProt ID: L8ITJ1, S7PCB4, F1S6I8, K7CS34, H2QBP0, D2GWM9, L5JP00, E2QSH5, S9WXI7, F1ME43, M1EPZ3, F7I5K2, I3ND78, M3Z8G4, H0XIF7, K7FR74, Q64731, Q8BQE0, G1U009, M0R6E1, G3RF46, G1SBX4, H2NRQ1, Q12952, Q498Y4). The GIPFL residues are part of the most conserved part of the C-terminal region.

Unfortunately, the proteins that had similar amino acid sequences to $FOXL1_{CTERM}$ did not have any published crystal structures. Since a suitable template protein was not available, structural information through homology modelling could not be acquired. In lieu of this, secondary structure bioinformatics tools were employed to predict $FOXL1_{CTERM}$ structure.

3.1.3. Secondary Structure Prediction

The secondary structure of FOXL1 was computationally predicted by PROFsec, a program found under the umbrella of the PredictProtein server.[33] The entire sequence was submitted to PROFsec, which suggested only the N-terminal domain possessed secondary structural elements. As shown in Figure 3.3b, PROFsec predicted that the N-terminal domain had four α-helical structure as well as two short β-strands. This prediction mostly agreed with the secondary structure elements published by Carlsson and Mahlapuu as shown in Figure 3.3a,[4] correctly predicting three α-helices and two β-strands. However there were a few discrepancies, including omitting one small β-strand, incorrectly predicting an α-helix, and both overestimating and underestimating the caps of helices and strands.

The C-terminal domain secondary structure elements were filtered and removed because this region had a much lower reliability index attributed to few diverse alignments in comparison to the N-terminal domain. In order to force a prediction of the C-terminal region, only the $FOXL1_{CTERM}$ sequence was used as input as shown in Figure 3.4b. With the new input sequence, PROFsec predicted that $FOXL1_{CTERM}$ had one short α-helix (residue 20-23) and four short β-strands (residue 34-38, 46-49; 53-56; 65-68). However, the prediction of each of the secondary structural elements was not reliable as indicated by a low reliability index scores as shown in Table 3.1. PROFsec assigns a structure reliability index score between 0 and 9 for each residue, where “0” means the prediction is not reliable while “9” shows high reliability. As seen in Table 3.1, the majority of structured residues had a reliability index of 3 or less. $FOXL1_{MUT}$ had a similarly predicted secondary structure, as shown in Figure 3.4a, where the major difference was that two β-strands had merged into a single β-strand due to the deletion of GIPFL. This prediction also had a low reliability index score.

Structure predictors can potentially be valuable tools to provide insight into protein secondary structure. However, in reference to FOXL1, although PROFsec worked fairly well for

the N-terminal domain, it did not provide any reliable structural information about the C-terminal domain. Since the secondary structure was not reliable for the study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$, it was not used as a starting structure for subsequent computer simulations.

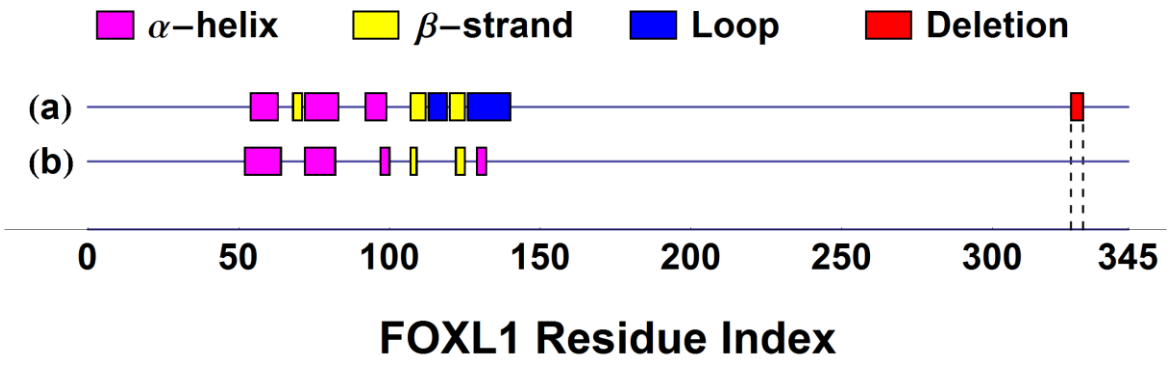


Figure 3.3: (a) The secondary structure elements of the N-terminal domain of FOXL1 has three α-helices spanning residue 54-63, 72-83, and 92-99; three β-strands from residue 68-71, 107-112, and 120-125; as well as two loops spanning residue 113-119 and 126-140.[14] (b) PROFsec secondary structure prediction of the entire sequence predicts the C-terminal domain to have no secondary structure elements while the N-terminal domain has four α-helical structure spanning residue 54-64, 72-82, 97-100, and 128-133, as well as two short β-strands from residues 108-109 and 20-125. There are a few discrepancies between the PROFsec predicted and published results, including omitting one small β-sheet (residue 68-71), incorrectly predicting an α-helix (residue 128-133), and being unreliable along the ends (caps) of helices and strands by both overestimating and underestimating the ends.

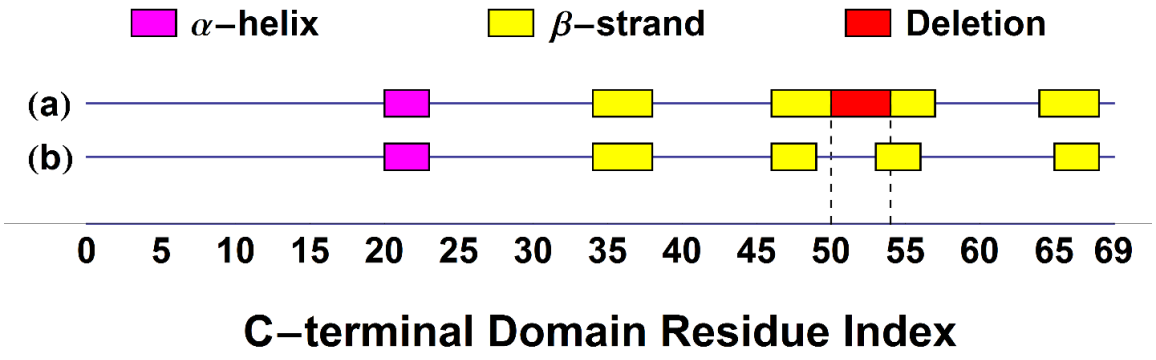


Figure 3.4: PROFsec secondary structure elements prediction of (a) $FOXL1_{MUT}$ and (b) $FOXL1_{CTERM}$. The deletion of GIPFL caused two β-strands $FOXL1_{CTERM}$ (residues from 46-49 and 53-56) to combine into a single β-strand in $FOXL1_{MUT}$ (residues from "46-57" with 50-54 removed).

Table 3.1: PROFsec prediction of secondary structure for $FOXL1_{CTERM}$. The prediction for each of the secondary structural elements was not reliable as indicated by a low reliability index score.

Structure	Residue Number	Reliability index (low = 0 ; high = 9)
Helix	21	1
	22	2
	23	1
	25	0
Extended	34	2
	35	3
	36	6
	37	5
	38	1
Extended	46	2
	47	3
	48	3
	49	1
Extended	53	1
	54	1
	55	0
	56	0
Extended	65	1
	66	2
	67	2
	68	0

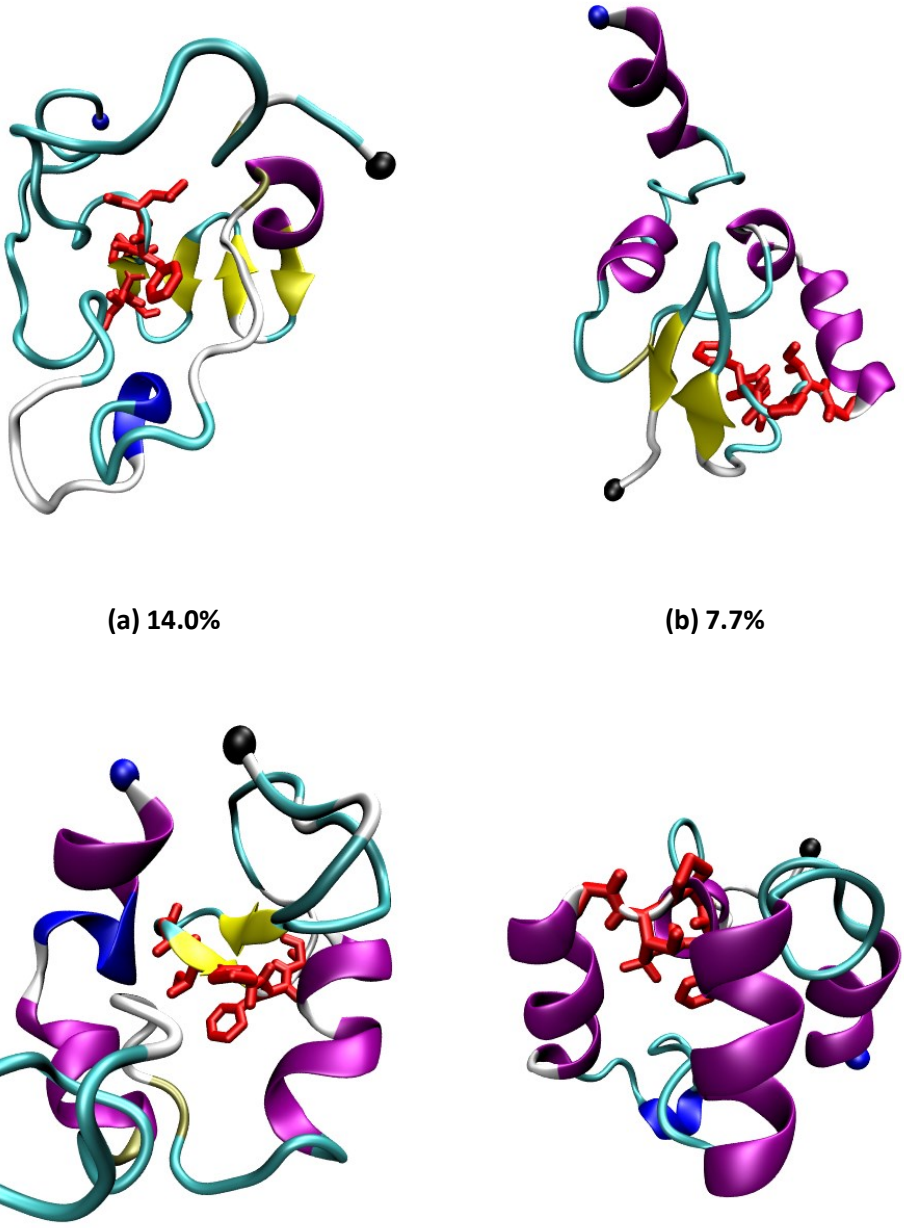
3.2. Computational

The tertiary structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ was computationally predicted using molecular dynamics (MD) simulations. The PACE force field was combined with replica exchange MD (REMD) in order to sample the folded configurational space of these proteins. The generated conformations for the 300 K replica were grouped according to the clustering scheme proposed by Heyer *et. al*.[55] using a 3 Å RMSD cut-off criteria between aligned C_α atoms.

3.2.1. $FOXL1_{CTERM}$

Representative $FOXL1_{CTERM}$ structures of the four statistically relevant clusters at 300 K are shown in Figure 3.5. These clusters had populations of 14.0%, 7.7%, 3.6%, and 1.6%. The top four cluster comprised about 26.9% of the total sampled structures. The other 73.1% of configurations

were part of much lower populated clusters or had very similar structures to these top clusters but fell outside the 3 Å cut-off. The highest populated structure (14.0%) of $FOXL1_{CTERM}$ had an N-terminal region that was mostly unstructured and a C-terminal region that had four short β-strands adjacent to a small α-helix. The GIPFL residues, indicated in red, were located in the core of the folded protein. In particular, the phenylalanine (F) sidechain was positioned directly towards the interior of the protein, which is to be expected considering phenylalanine is a hydrophobic amino acid. The second most populated $FOXL1_{CTERM}$ structure (7.7%) was more structured than the first cluster, since it had four N-terminal α-helices and two short C-terminal β-strands. Again, the GIPFL residues were found in the core of the folded protein. The third most populated structure (3.6%) was similar to the second and had three N-terminal α-helices and two short C-terminal β-strands. The glycine and isoleucine of the GIPFL segment formed one of the β-strands. The GIPFL residues were again in the interior of the folded protein. Finally, the fourth most populated $FOXL1_{CTERM}$ structure (1.6%) formed a three-helix bundle with the GIPFL sidechains contributing to the hydrophobic core of the domain. The GIPFL sequence was located in a helix packed on top of the helix bundle.



(a) 14.0% **(b) 7.7%**

(c) 3.6% **(d) 1.6%**

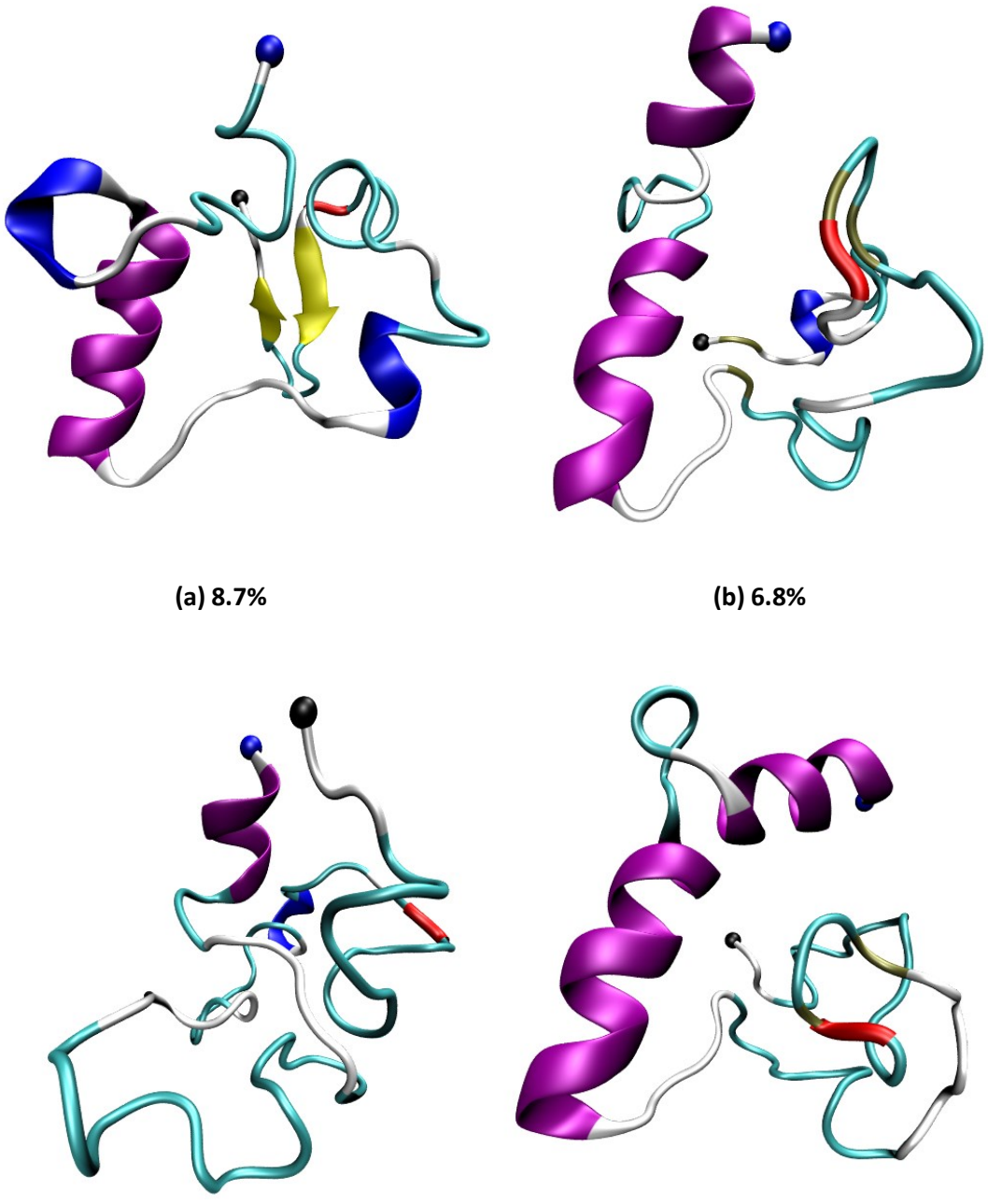
Figure 3.5: Representative structures of the four most populated clusters for FOXL1$_{CTERM}$ with their respective populations indicated below the structures. The population is the percentage of the total samples structures from the full 4000 ns simulation that are within a 3 Å RMSD criteria of the represented structure. The blue and black spheres mark the N-terminus and C-terminus, respectively. The GIPFL residues have been indicated in red.

3.2.2. $FOXL1_{MUT}$

Figure 3.6 shows representative $FOXL1_{MUT}$ structures of the four largest clusters at 300 K. These clusters had populations of 8.7%, 6.8%, 5.0%, and 4.4%. The highest populated structure (8.7%) had an N-terminal region with a single α-helix and a C-terminal region with two short β-strands. The GIPFL deletion was indicated in red. With the GIPFL residues missing, the surrounding region was now solvent exposed on the surface of the protein. The second most populated $FOXL1_{MUT}$ structure (6.8%) was less structured than the first, having two N-terminal α-helices and disordered C-terminal region. Again, due to the GIPFL deletion, the surrounding region was on the exterior on the protein. The third most populated structure (5.0%) was completely unstructured except for a small N-terminal α-helix. Finally, the fourth most populated $FOXL1_{CTERM}$ structure (4.4%) was similar to the second cluster and had two N-terminal α-helices and disordered C-terminal region.

3.2.3. Comparison

The computationally predicted clusters of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ were completely different. First of all, the $FOXL1_{CTERM}$ clusters were more structured and folded than $FOXL1_{MUT}$. The $FOXL1_{CTERM}$ clusters had a significant number of α-helices and β-strands while the mutant generally possessed one or two α-helices. Secondly, the GIPFL residues in $FOXL1_{CTERM}$ are involved in structural elements (i.e., a β-strand for the first and third clusters, and an α-helix for the fourth) or directly adjacent to structure (i.e. second cluster). In contrast, the $FOXL1_{MUT}$ clusters were significantly disordered in the C-terminal region containing the deletion, with the exception of the first cluster that has two short β-strands adjacent the deletion. In the $FOXL1_{CTERM}$ clusters, the GIPFL residues are located in the hydrophobic core, with the phenylalanine sidechain positioned towards the interior of the protein. These computational results suggested that the deletion of GIPFL residues in the $FOXL1_{MUT}$ system disrupted the hydrophobic core as well as particular structural elements, causing the region adjacent the mutation to become randomly coiled. From these results, we hypothesize that the mutation alters the structure of the protein-protein binding surface and thereby hinders the C-terminal domain from binding effectively to co-regulatory protein(s), preventing FOXL1 from carrying out its regulatory functions.



(a) 8.7% **(b) 6.8%**

(c) 5.0% **(d) 4.4%**

Figure 3.6: Representative structures of the four most populated clusters for FOXL1$_{MUT}$ with their respective populations indicated below the structures. The population is the percentage of the total samples structures from the full 4000 ns simulation that are within a 3 Å RMSD criteria of the represented structure. The blue and black spheres marks the N-terminus and C-terminus, respectively. The location the GIPFL deletion is indicated in red.

3.3. Experimental Results

3.3.1. Introduction

This section details the progress made towards expressing, purifying, and structurally characterizing $FOXL1_{CTERM/MUT}$ using three different constructs: (1) 6His—$FOXL1_{CTERM/MUT}$, (2) S-tag—6HIS—$FOXL1_{CTERM}$, and (3) SN—$FOXL1_{CTERM/MUT}$—6HIS. The first construct investigated, 6His-$FOXL1_{CTERM/MUT}$, had low protein expression and significant sample impurity, which hindered identification and purification of the target protein. The second construct studied, S-tag—6HIS—$FOXL1_{CTERM}$, also had low protein expression and sample impurities, but allowed for identification of the target protein by western blot and enabled structural characterization of an impure target protein by circular dichroism. The final construct, SN— $FOXL1_{CTERM/MUT}$—6HIS, had high expression yields, was easily identified via western blot, was successfully enriched, and allowed for some structural characterization of the target protein by circular dichroism. The following sections discuss the challenges and successes for each construct expression in relevant detail.

3.3.2. Construct 1: 6His—$FOXL1_{CTERM/MUT}$

The 6His—$FOXL1_{CTERM/MUT}$ construct was expressed six times (four with $FOXL1_{CTERM}$—6His and two with $FOXL1_{MUT}$—6His) following the protocol in the Methodology section. The desired protein was assumed to be present in inclusion bodies based on preliminary data from a co-worker.[‡‡] This assumption could not be confirmed experimentally due to difficulties in identification and purification.

Identification and purification of the target protein was complicated due to low protein expression and significant sample impurity. This is apparent in Figure **3.7**, which shows a silver stained gel of $FOXL1_{CTERM}$—6His (~8.0 kDA) after Ni column purification that contained four major protein bands at approximately 8.5, 10, 17, and 24 kDa. This gel required the highly sensitive silver staining, which had a detection limit of 1 ng, in order to detect protein bands. In most cases, Coomassie staining of a gel, with a detection limit of 100 ng, was not able to detect any protein indicating low overall protein expression. Low or unsuccessful protein expression was further

[‡‡] Dr. Ahmed Mostafa, Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada

reinforced by a spotless His-tag western blot, which had a sensitivity range between from 4 ug/mL to 1 ng/mL. Apart from a silver stained gel that showed a band with the expected molecular weight, the presence of FOXL1 $_{CTERM/MUT}$ was not confirmed by any other method, including MALDI (due to high salt or low protein concentration), western blot, and mass spectrometry.

Low protein expression can be caused by a number of complex factors. One possible cause is that the foreign protein may be toxic to the host strain. If an expressed protein is toxic to the host, the proteins may be degraded by the host cell or the host can undergo lysis. Other causes for low or no protein expression in a host protein include initiation problems and differences from the source organism in environment, chaperones, and codon use bias.[88]

Enrichment of the desired protein was also difficult because the protein eluted early in the Ni affinity purification, specifically between 10 – 50 mM of imidazole, and consequently co-eluted with at least three other major impurities. Further purification was attempted using HPLC, which failed due to insolubility of product in the solvent. Purification using water solubility differences was also tried but ineffective because the desired protein and impurities all precipitated in water. This last purification attempted was size exclusion gel filtration, which was not effective because the protein in eluted fractions were too dilute to detect by UV-vis or silver stain gel electrophoresis. The difficult purification and identification of the target protein motivated the development of a new protein construct.

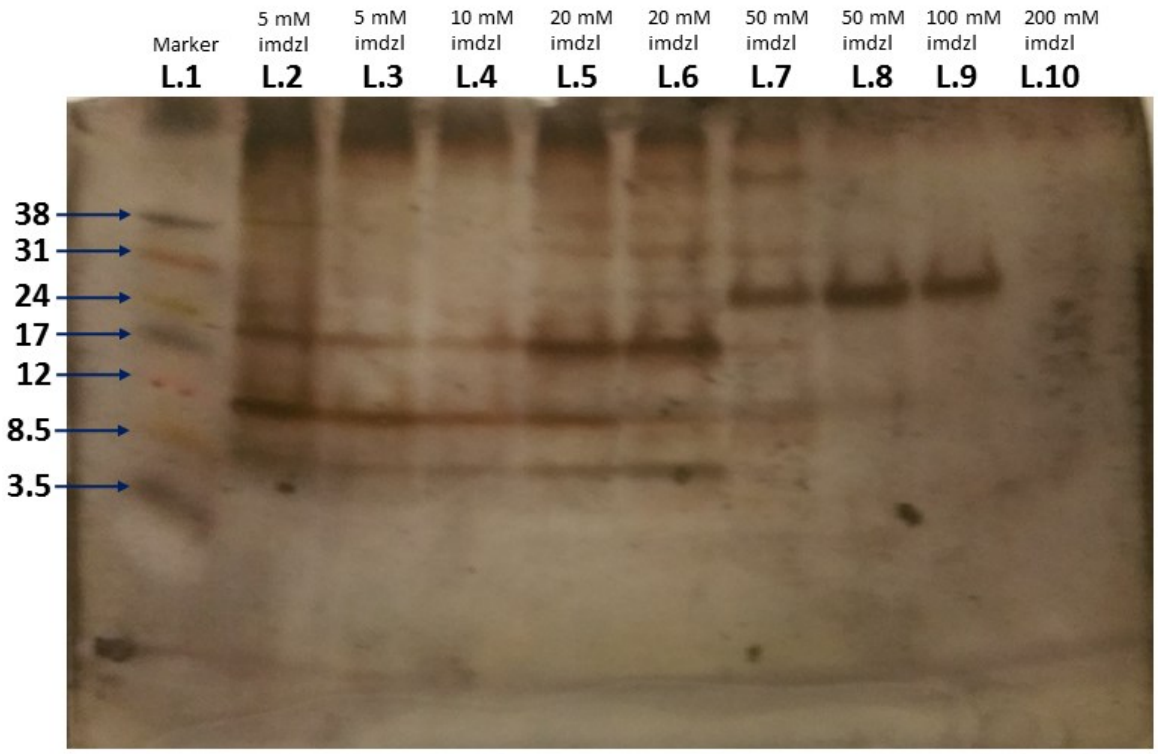


Figure 3.7**: Silver stained 16.5% tris-tricine gel of FOXL1$_{CTERM}$—6His (~8.0 kDa) following Ni column purification using elution scheme 1 in Table 2.1. Four major bands are observed at 8.5 kDa, 10 kDa, 17 kDa, and 24 kDa that significantly overlap during elution. The protein eluted between 10 – 50 mM of imidazole as seen in L.2 – L.6.**

3.3.3. Construct 2: S-tag—6His—FOXL1$_{CTERM}$

The S-tag—6HIS—FOXL1 $_{CTERM}$ construct was expressed three times using the protocol detailed in the Methodology section. In order to determine if the target protein was successfully expressed and whether it remained soluble, associated with the membrane, or formed inclusion bodies, a gel and S-tagged western blot was run in duplicate on these three crude fractions. As seen in Figure 3.8, the Coomassie stained gel did not reveal well-defined protein bands, which is a common problem for crude protein mixtures. On the other hand, the western blot showed that the majority of S-tagged protein formed inclusion bodies (L.5), while only some associated with the cellular membrane (L.4), and very little remained soluble (L.3). Unfortunately, there were two S-tagged protein in the inclusion bodies having approximately molecular weights between 8.5-12 kDa and 12-24 kDa. As demonstrated later in other gels, these two bands fluctuate relative to the reference marker, thereby making assignment of the molecular weight difficult. The calculated molecular weight of S-tag—6HIS—FOXL1$_{CTERM}$ was 11.2 kDa, and thus it was unclear which of the two bands belongs to the target protein. The higher band could not be explained by dimerization since this construct did not contain any cysteine residues to form inter-protein disulfide bonds. Protein aggregation was also ruled out since the initial gel was run under denaturing (SDS-PAGE) conditions. One possibility that accounted for the second S-tagged protein band was that transcription started or ended incorrectly resulting in a transcriptional artifact. In contrast to the inclusion body fractions, the soluble fraction contained small amounts of only the higher molecular weight band. Thus, further investigation was required for the supernatant and the inclusion bodies (or combined membrane/ inclusion body) fractions.

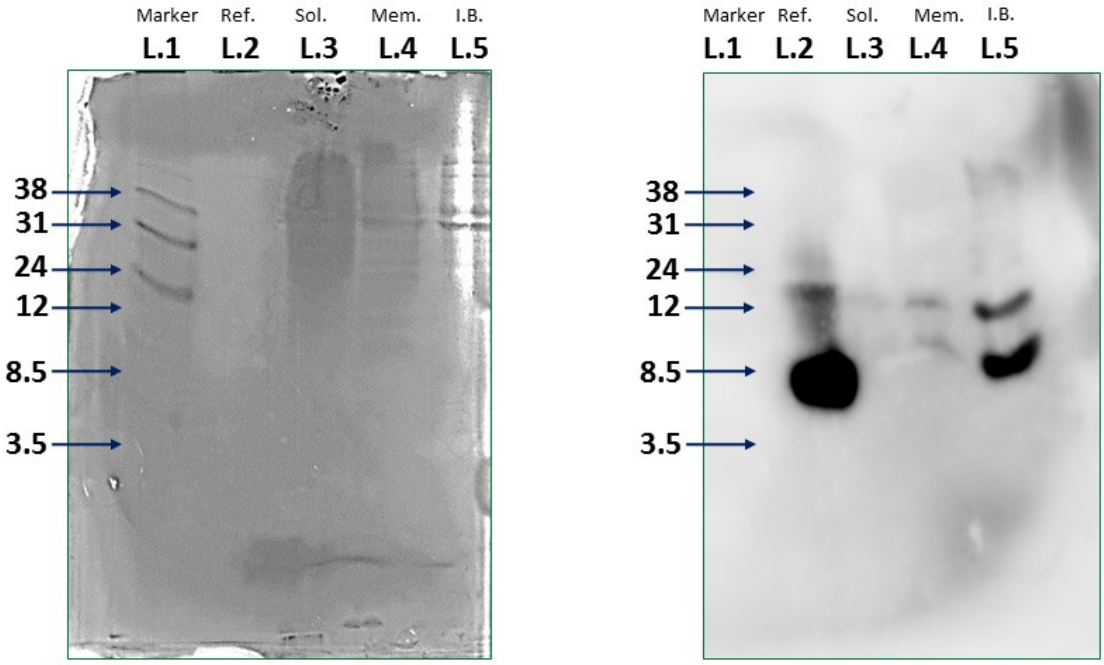


Figure 3.8: Coomassie stained 16.5% tris-tricine gel (left) and S-tag western blot (right) of the crude fractions after cell lysis to determine if the S-tag—6HIS—FOXL1 $_{CTERM}$ (~11.2 kDa) target protein remained soluble (L.3), associated with the membrane (L.4), or formed inclusion bodies (L.5). An S-tagged protein (L.2) was used as a reference to prove the western blot worked. (Left) The Coomassie stained gel showed blurred bands, which is common for crude samples. (Right) The western blot revealed two S-tagged proteins present in inclusions bodies and associated with the membrane, with the majority of protein in inclusion bodies. These proteins had a molecular weight of 8.5-12 kDa and 12-24 kDa as measured by comparison to the marker (L.1.). It is unclear which band was the target protein. The soluble fraction appeared to only have a single, faint band between 12-24 kDa.

3.3.3.1. Inclusion Bodies Sample

The crude, solubilized inclusion body sample (L.5 in Figure 3.8) was enriched by first using DE52 to remove bacterial cell impurities, followed by Ni affinity chromatography using elution scheme 4 detailed in Table 2.1. Coomassie stained gel and S-tag western blot of the Ni column purification is depicted in Figure 3.9. The western blot revealed that two S-tagged proteins with approximate molecular weights of 8.5-12 kDa and 12-24 kDa co-eluted at 100 mM of imidazole (L.8). The corresponding Coomassie stained gel showed no bands in L.8, signifying that the loaded sample had less than 100 ng of protein and, overall, these S-tagged proteins had low expression yields. Another gel and S-tag western blot was run on consecutive fractions eluted with 100 mM of imidazole as shown in Figure 3.10. (For reference, L.8 from Figure 3.9 represented the same sample as L.6 in Figure 3.10.) The silver stained gel showed that these fractions were impure and contained at least three proteins with a molecular weight of 8.5-12 kDa, 12-24 kDa, and 31 kDa. The corresponding western blot again revealed two S-tagged proteins between 8.5-12 kDa and 12-24 kDa. The absorbance of the fractions represented by L.2 to L.10 in Figure 3.10 was less than 0.003 at 280 nm as measured against a blank containing 100 mM imidazole, 0.2% CHAPS, 6 M urea, and TBS. The low absorbance of these fractions, combined with the inability to detect protein via Coomassie stain indicated low protein expression.

Fractions represented by L.1 to L.5 of Figure 3.10 was prepared for structural determination by circular dichroism. The sample was dialyzed against distilled water to remove the salt and imidazole, and then freeze dried to yield in 1 mg of white, solid product. The 1 mg of product was dissolved in 500 µL of distilled water. The resultant sample had an absorbance of 0.123 at 280 nm, which equated to approximately 48 µM of protein. The circular dichroism spectrum of this sample is shown in Figure 3.11. The CD spectrum exhibited low overall signal-to-noise and substantial fluctuations at the lower wavelength - consistent with a lower than ideal protein concentration with some small molecular contaminants that scatter at the smaller wavelengths. However, even with these shortcomings it was still possible to observe minima at 208 and 222 nm suggesting an overall α-helical structure. Further enrichment of this sample was not attempted due to the very low concentration of protein.

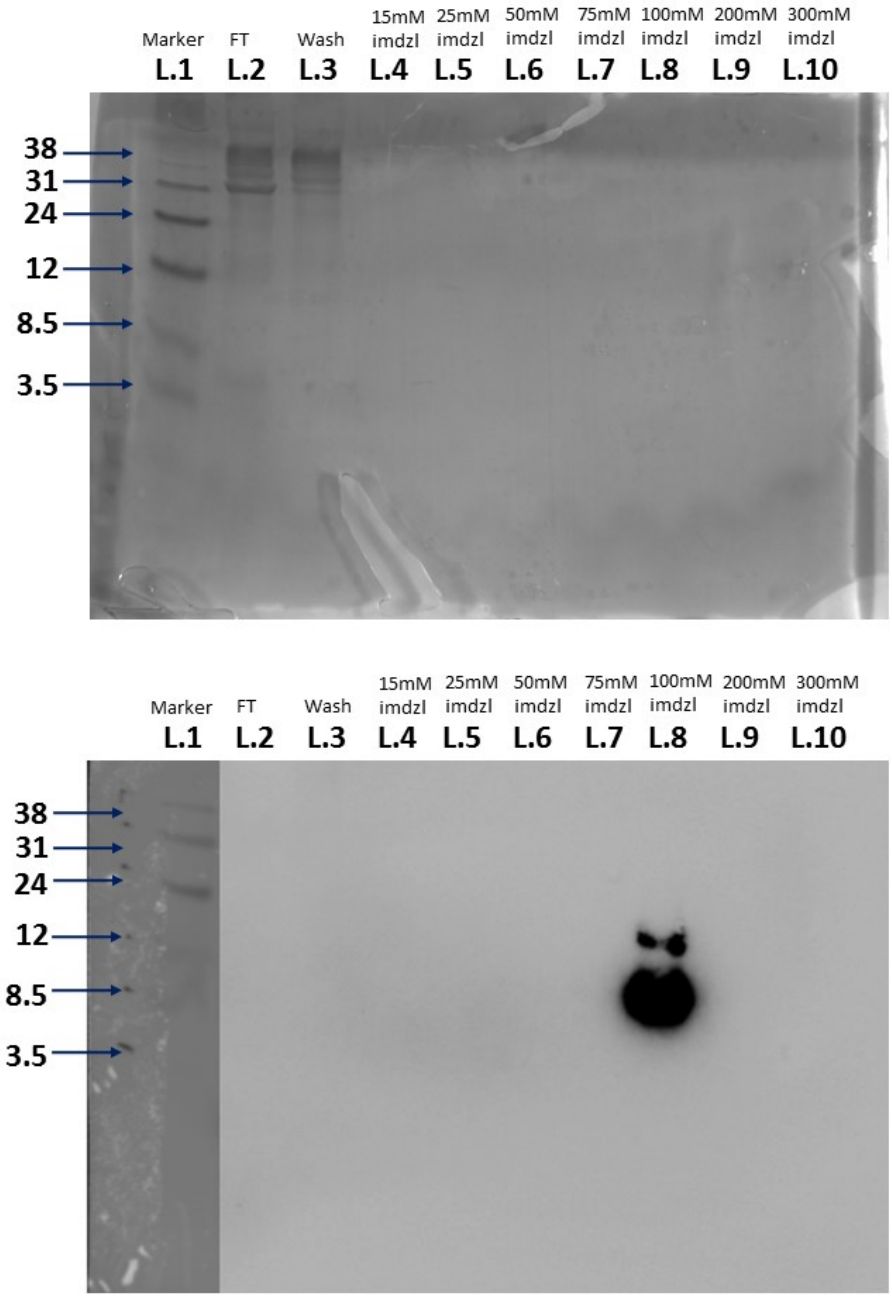


Figure 3.9: Coomassie stained 16.5% tris-tricine gel (top) and S-tag western blot (bottom) after Ni column purification of the inclusion bodies fraction (see Figure 3.8, L.5) to find S-tag—6HIS—FOXL1 $_{CTERM}$ (~11.2 kDa). Protein was eluted using elution scheme 4 detailed in Table 2.1. Each figure shows a reference marker (L.1), flow-through (L.2), wash (L.3), and select fractions eluted using various imidazole concentrations (L.4 – L.10). (Top) Gel did not reveal any bands in L.4-L.10, suggesting that any protein present was below the detection limit of 100 ng. (Bottom) Western blot revealed that two S-tagged proteins, with molecular weights between 8.5-12 kDa and 12-24 kDa, co-eluted at 100 mM of imidazole (L.8).

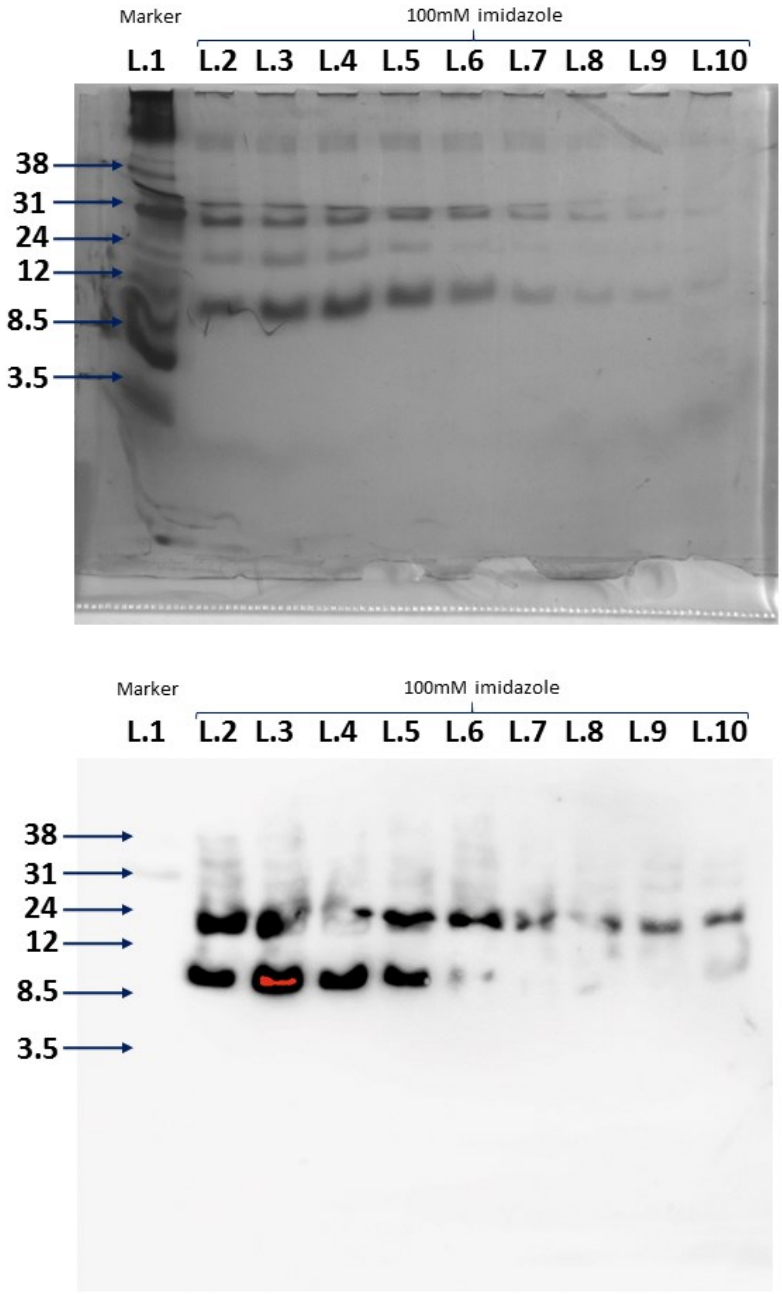


Figure 3.10: Silver stained 16.5% tris-tricine gel (top) and S-tag western blot (bottom) after Ni column purification of the inclusion bodies fraction showing consecutive fractions eluted using 100 mM of imidazole to find S-tag—6HIS—FOXL1$_{CTERM}$ (~11.2 kDa). (Top) Gel showed the sample was impure, as three protein bands at 8.5-12 kDa, 12-24 kDa, and 31 kDa were present. (Bottom) Western blot revealed that two S-tagged proteins, with approximate molecular weights between 8.5-12 kDa and 12-24 kDa co-eluted at 100 mM of imidazole.

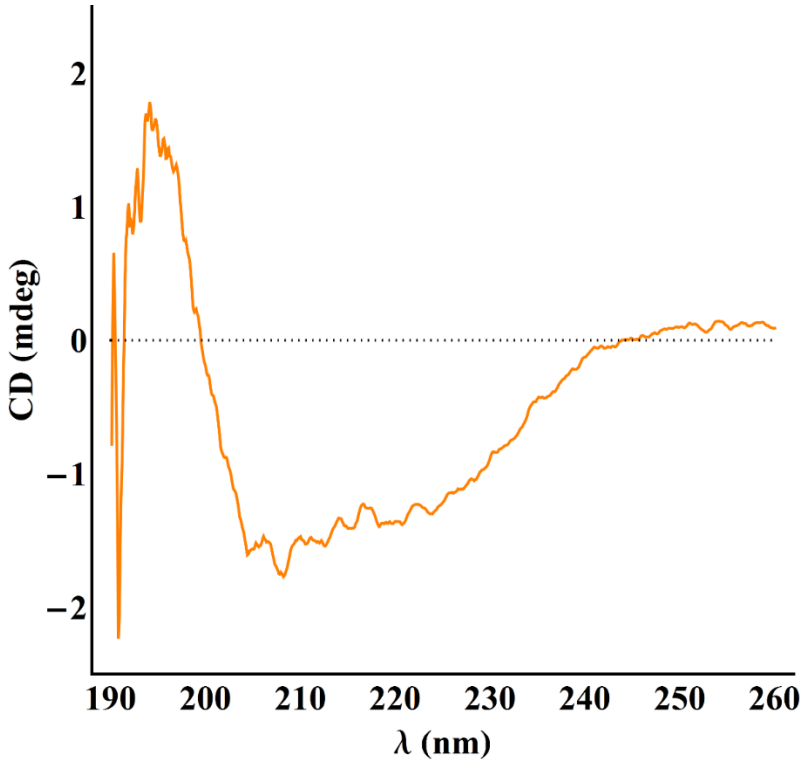


Figure 3.11: Far UV CD spectra of an impure sample of S-tag—6HIS—FOXL1$_{CTERM}$ (~48 µM) in distilled water with a 0.5 mm quartz cuvette at room temperature, where 20 scans were averaged. The CD spectrum of FOXL1$_{CTERM}$ showed the characteristic features of a predominant helical protein revealed by minima at 208 nm and 212 nm.

3.3.3.2. Soluble Sample

The crude, soluble sample (L.3 in Figure 3.8) was enriched by first using DE52 to remove bacterial cell impurities, followed by Ni affinity chromatography using elution scheme 3 detailed in Table 2.1. A silver stained gel and S-tag western blot of the Ni column purification is depicted in Figure 3.12. The gel showed three protein with approximate weights of 8.5-12, 24-31, and >38 kDa were present. The corresponding western blot revealed that two S-tagged proteins with molecular weights of 8.5-12 and >38 kDa that co-eluted at 50 mM of imidazole (L.2-L.10 on gel).

Fractions represented by L.5 to L.10 of Figure 3.12 were prepared for structural determination by circular dichroism. The sample was dialyzed against distilled water to remove the salt and imidazole, which resulted a white solid precipitating out of solution. This solid was re-dissolved by lowering the pH from 5.6 to 3.2 using HCl, then bringing it back up to a pH of 3.9 using NaOH. The absorbance of this sample was 0.311 at 280 nm which equated to 121 µM of protein. The circular dichroism spectrum of this sample, shown in Figure 3.13, exhibited minima at 208 and 222 nm which suggested a dominant helical shape for the protein(s) in solution. Similar to the inclusion body sample, it was possible that S-tag—6HIS—FOXL1$_{CTERM}$ had a helical secondary structure, but due significant sample impurity this was not confirmed. Purification of this sample was not attempted due to low protein concentration.

This construct was problematic because of low protein expression and difficulties in identifying the target protein due to the presence of two expressed S-tagged proteins. This motivated the investigation of a new protein construct.

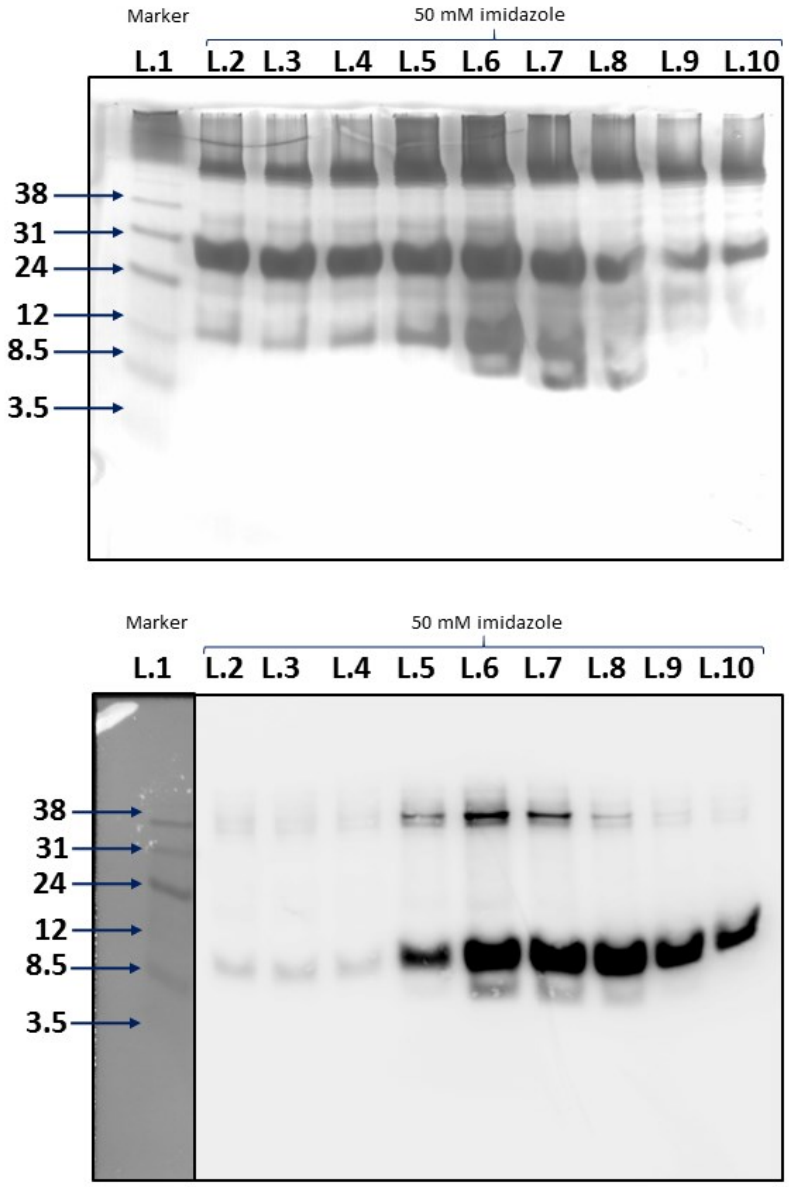


Figure 3.12: Silver stained 16.5% tris-tricine gel (top) and S-tag western blot (bottom) after Ni column purification of the soluble fraction showing consecutive fractions eluted using 50 mM of imidazole to find S-tag—6HIS—FOXL1$_{CTERM}$ (~11.2 kDa). (Top) Gel showed the sample was impure, as three significant protein bands at 8.5-12, 24-31, and >38 kDa were present. (Bottom) Western blot revealed that two S-tagged proteins, with approximate molecular weights between 8.5-12 kDa and >38 kDa co-eluted at 50 mM of imidazole.

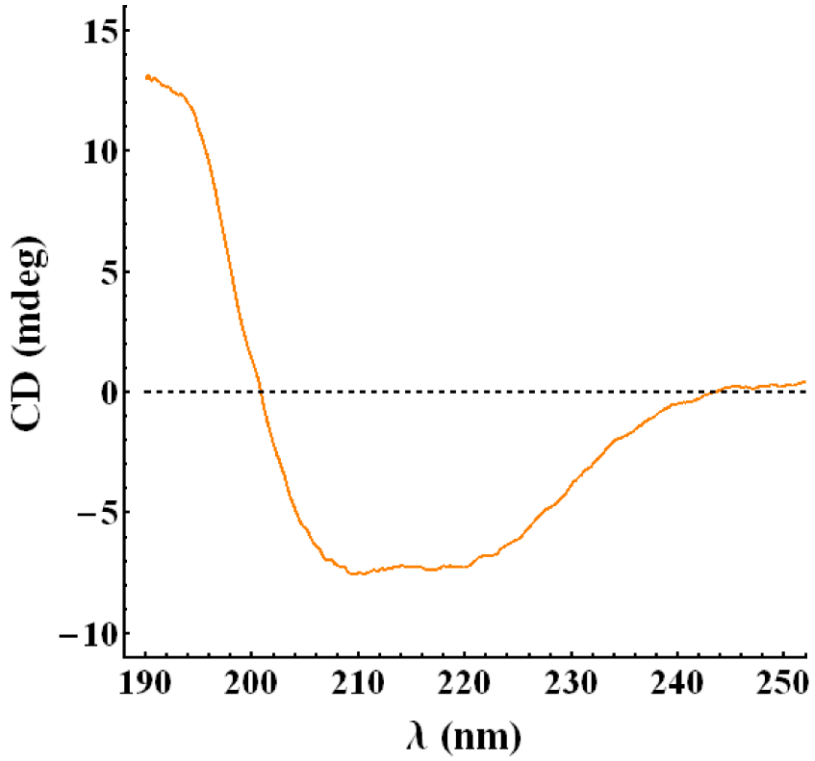


Figure 3.13: Far UV CD spectra of an impure sample of S-tag—6HIS—FOXL1$_{CTERM}$ (121 µM) in distilled water at pH 3.9 with a 0.5 mm quartz cuvette at room temperature, where 5 scans were averaged. The CD spectrum showed the characteristic features of a predominant helical protein, revealed by minima at 208 nm and 212 nm.

3.3.4. Construct 3: SN fusion—FOXL1$_{CTERM/MUT}$—6His

The next construct investigated was designed with an N-terminal SN tag to try and increase the protein expression levels as well as a C-terminal 6HIS tag to facilitate purification. The SN—FOXL1$_{CTERM/MUT}$—6HIS construct was expressed five times (three times with SN—FOXL1$_{CTERM}$—6HIS and twice with SN—FOXL1$_{MUT}$—6HIS) using the protocol detailed in the Methodology section. In order to determine if the target protein remained soluble, associated with the membrane, or formed inclusion bodies, Ni affinity column purification was run on each of the crude samples for SN—FOXL1$_{CTERM}$—6HIS and SN—FOXL1$_{MUT}$—6HIS. In both cases, the target protein primarily formed inclusion bodies upon expression. This was illustrated by the Coomassie stained gel in Figure 3.14 for SN—FOXL1$_{CTERM}$—6HIS (~26.0 kDa) and in Figure 3.15 for SN—FOXL1$_{MUT}$—6HIS (~25.5 kDa), which showed a very large band at ~26 kDa that was eluted with 100 mM imidazole. A western blot probing for the SN moiety confirmed that SN—FOXL1$_{CTERM}$—6HIS was successfully expressed in inclusion bodies which is shown in Figure 3.16 by the large blot at ~26 kDa in L.5-L.10. It was evident that this construct worked significantly better than the first two constructs tested due to its (1) significantly higher protein expression and (2) the higher imidazole concentration required for target protein elution during Ni affinity purification.

In order to prevent the α-helical SN-fusion protein from influencing the structure of FOXL1$_{CTERM/MUT}$, a CNBr digest was employed to cleave SN off the expressed protein. To optimize the reaction time for CNBr digest, a Coomassie stained gel was run that monitored the CNBr digest over time for SN—FOXL1$_{CTERM}$—6HIS (refer to Section 2.2.2.12 for process details) which is shown in Figure 3.17. At time zero (L.2), the expressed protein (SN—FOXL1$_{CTERM}$—6HIS) showed a large band at ~26 kDa. After 16 hours (L.3), the starting protein was partially digested into the SN-fusion protein (17.9 kDa) and FOXL1$_{CTERM}$—6HIS (8.1 kDa). The amount of FOXL1$_{CTERM}$—6HIS protein did not appear to increase after 24 hours (L.4). However, after 40 hours (L.6), another species resulted that was smaller in size than FOXL1$_{CTERM}$—6HIS. This suggested the CNBr chemical was instead digesting the SN or FOXL1$_{CTERM}$—6HIS protein. The expressed protein was never fully digested, even after 3 days. Thus, the optimal time for CNBr digest was between 24-32 hours to obtain about 50% digestion of SN—FOXL1$_{CTERM}$—6HIS into SN-fusion protein and FOXL1$_{CTERM}$—6HIS. A similar result was observed for the mutant protein.

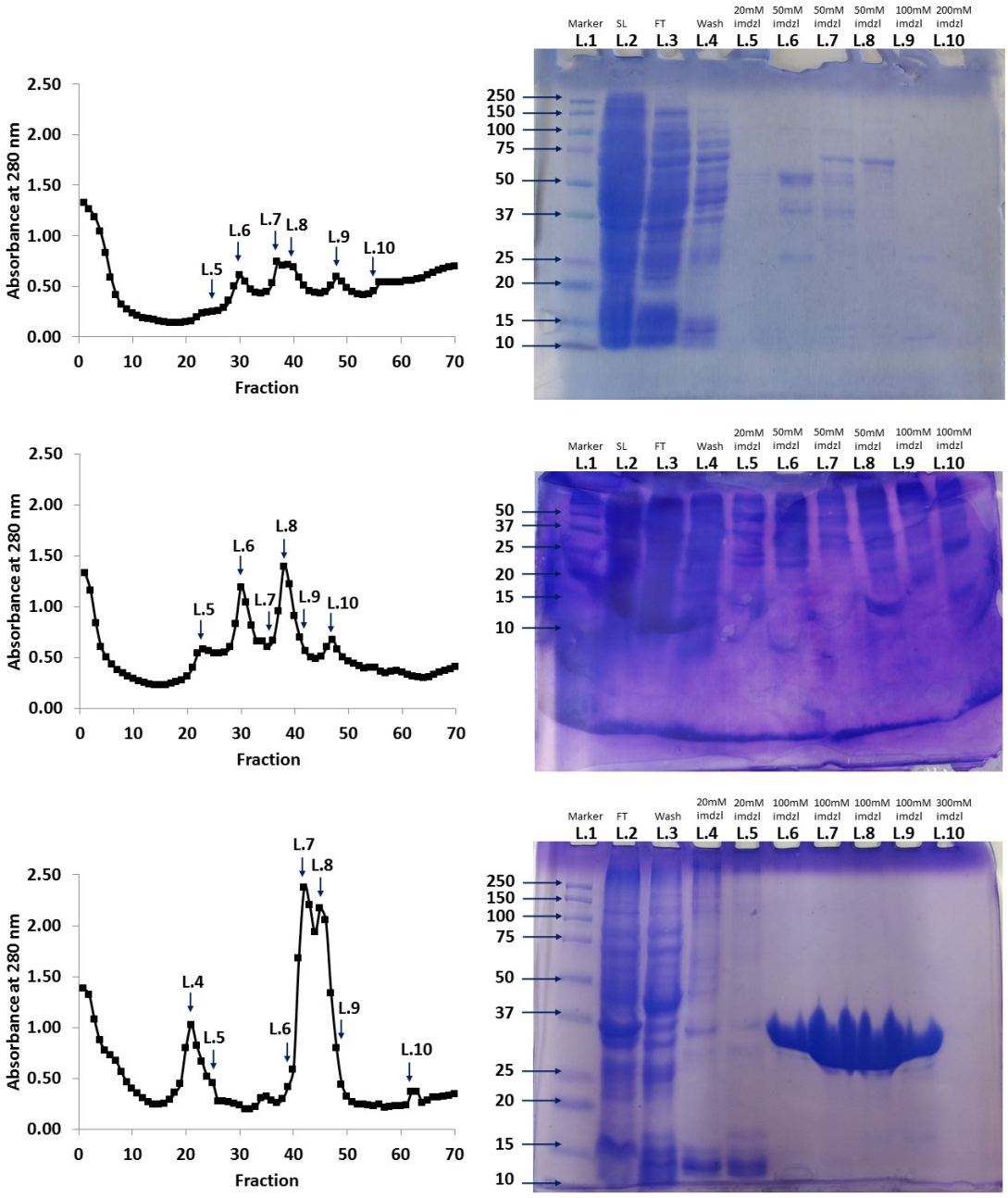


Figure 3.14 UV-vis absorbance (left) and Coomassie stained gel (right) of the fractions following Ni affinity column purification to determine if the SN—FOXL1$_{CTERM}$—6HIS (~26.0 kDa) protein remained soluble (top, 12% polyacrylamide tris-glycine gel), associated with the membrane (middle, 16.5% tris-tricine gel), or formed inclusion bodies (bottom, 12% polyacrylamide tris-glycine gel). Elution scheme 3, 8, and 2 detailed in Table 2.1 was employed for Ni column chromatography of the top, middle, and bottom experiments, respectively. The Coomassie stained gel revealed that significant amounts of protein with the expected molecular weight of N—FOXL1$_{CTERM}$—6HIS expressed as inclusion bodies (L.6-L.9, bottom) and required at least 100 mM of imidazole for elution.

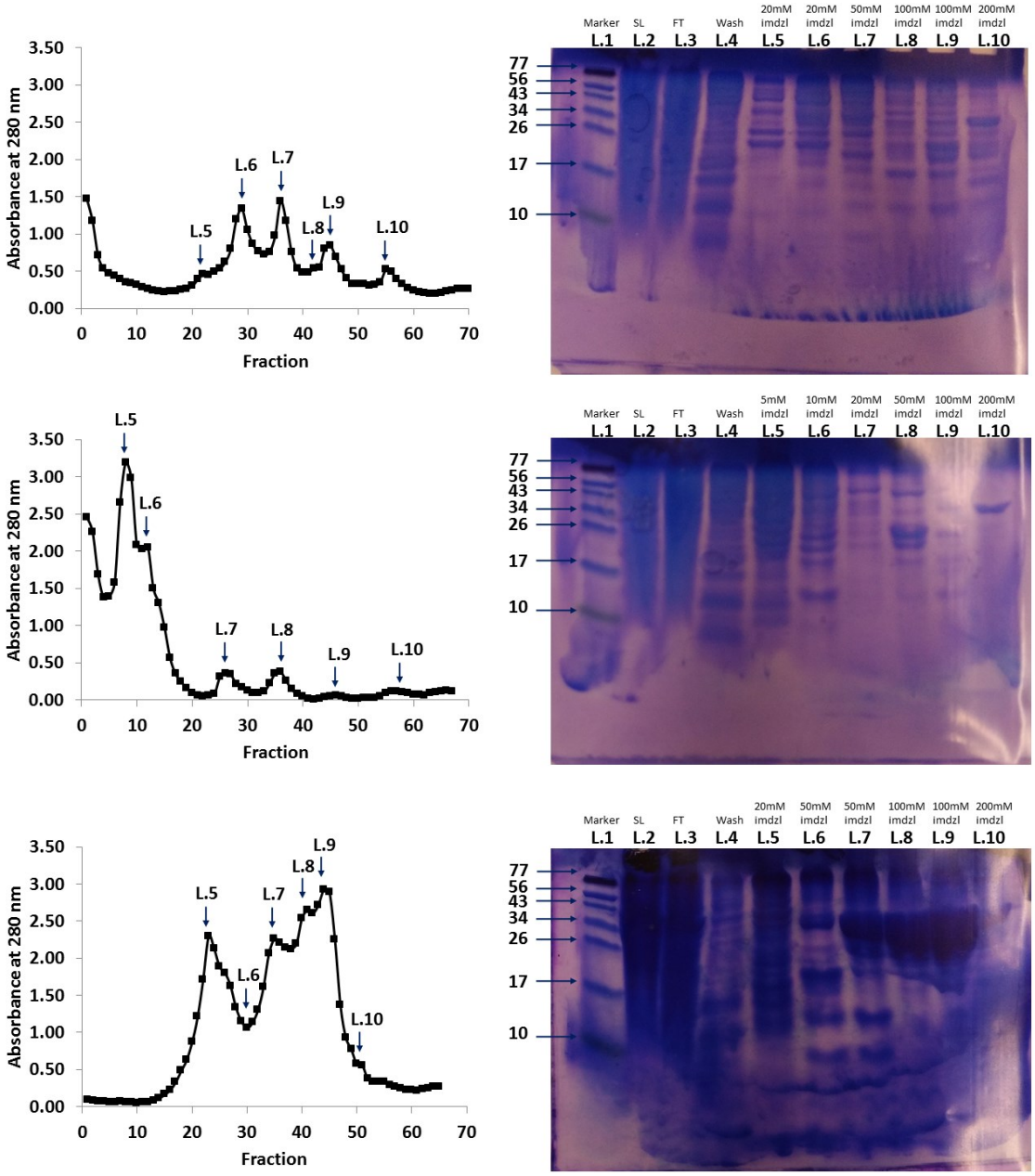


Figure 3.15: UV-vis absorbance (left) and Coomassie stained 16.5% tris-tricine gel (right) of the fractions following Ni affinity column purification to determine if the SN—FOXL1$_{MUT}$—6HIS (~25.5 kDa) protein remained soluble (top), associated with the membrane (middle), or formed inclusion bodies (bottom). Elution scheme 3, 8, and 2 detailed in Table 2.1 were employed for Ni column chromatography of the top, middle, and bottom experiments, respectively. The Coomassie stained gel revealed that significant amounts of protein with the expected molecular weight of SN—FOXL1$_{MUT}$—6HIS expressed as inclusion bodies (L.7-L.10) and required at least 100 mM of imidazole for Ni column elution.

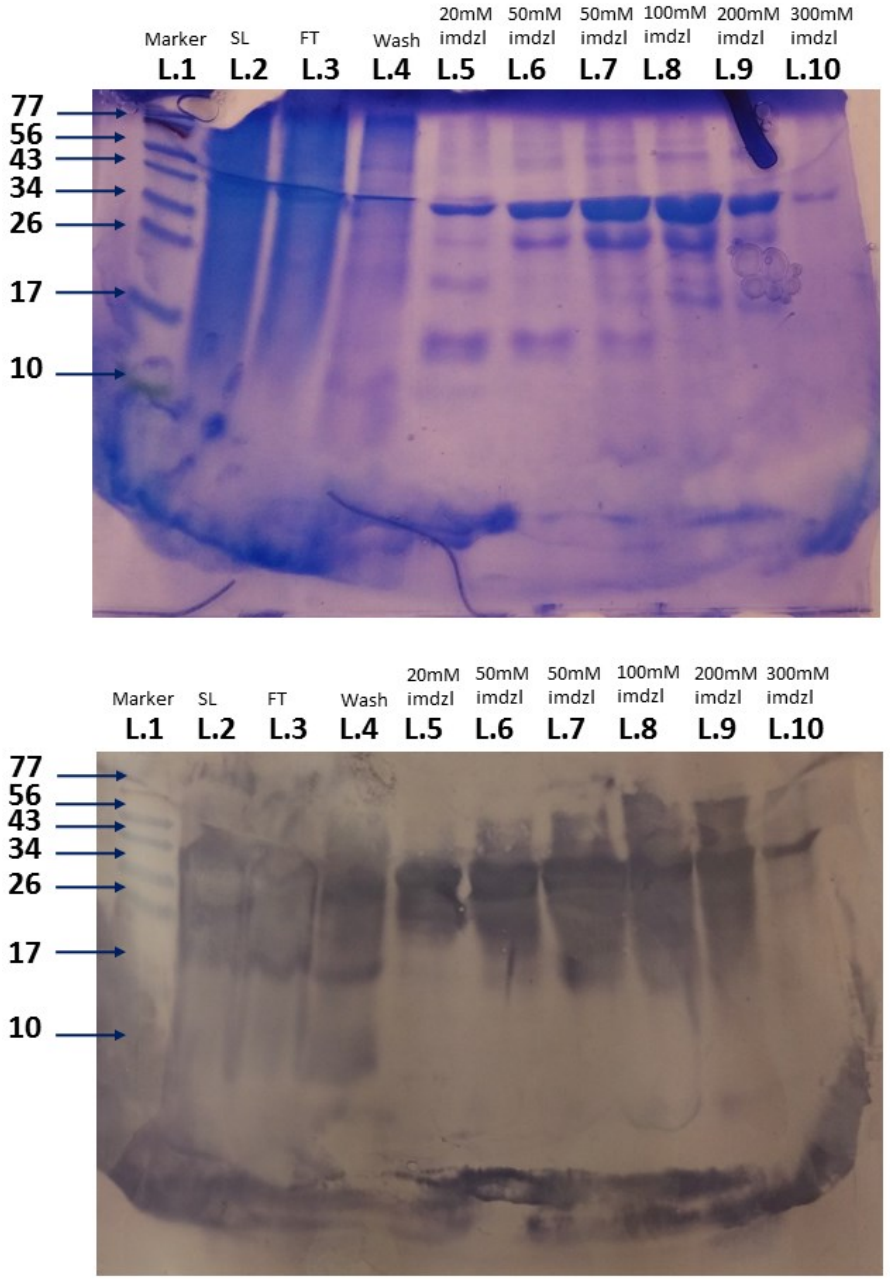


Figure 3.16: Coomassie stained 16.5% tris-tricine gel (top) and SN-tag western blot (bottom) after Ni column purification of the inclusion bodies sample to find SN—FOXL1$_{CTERM}$—6HIS (~26.0 kDa). Protein was eluted using elusion scheme 2 detailed in Table 2.1. Each figure shows a reference ladder (L.1), the sample loaded (L.2), flow-through (L.3), wash (L.4), and select fractions eluted using various imidazole concentrations (L.4 – L.10). (Left) The Coomassie stained gel revealed that significant amounts of protein with the expected molecular weight of SN—FOXL1$_{CTERM}$—6HIS expressed as inclusion bodies and required at least 100 mM of imidazole for Ni column elution. (Right) An SN western blot confirmed that the target protein, with approximate molecular weights of 26.0 kDa, eluted at 100 mM of imidazole (L.5-L.10).

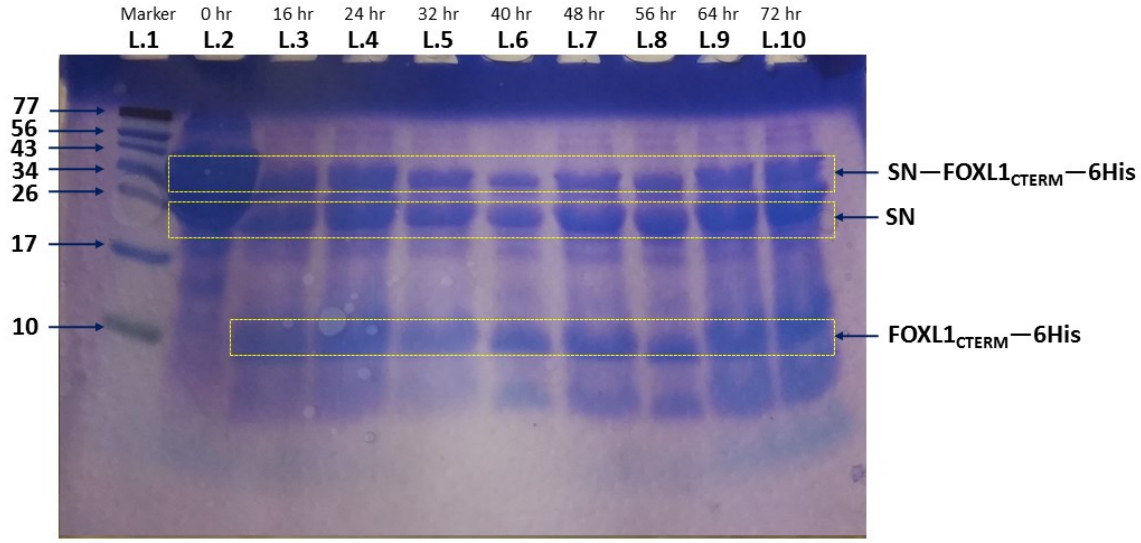


Figure 3.17: Coomassie stained 16.5% tris-tricine gel showing the CNBr digest of SN—FOXL1$_{CTERM}$—6HIS (~26.0 kDa) into SN-tag (17.9 kDa) and FOXL1$_{CTERM}$—6HIS (~8.5 kDa) over three days. The amount of FOXL1$_{CTERM}$-6HIS protein did not appear to increase after 24 hours (L.4). At 40 hours (L.6), another unknown impurity resulted that was smaller in size than FOXL1$_{CTERM}$—6HIS. The optimized CNBr digest time was between 24-32 hours to obtain 50% digestion of SN—FOXL1$_{CTERM}$—6HIS into SN-fusion protein and FOXL1$_{CTERM}$—6HIS.

3.3.4.1. Sample Purification

After CNBr digest, the target proteins (FOXL1$_{CTERM/MUT}$—6HIS) now possessed at least two other impurities including the expressed protein (SN—FOXL1$_{CTERM/MUT}$—6HIS) and SN-fusion protein. Therefore, further purification was required to isolate the target protein. Numerous purification protocols were tested to isolate the target protein from the expressed protein and SN-fusion protein. These include experiments to separate the proteins based on different Ni binding affinities and molecular weights. Overall, the target protein was successfully isolated from the SN-fusion protein and the expressed protein using Ni affinity chromatography and size exclusion gel filtration, respectively, as discussed below.

First of all, since the SN-fusion protein was no longer bound to a His-tag, it was easily removed using Ni affinity column purification using elution scheme 2 detailed in Table 2.1. As shown in Figure 3.18, the SN-fusion protein eluted in the flow-through (L.3) and wash (L.4), while SN—

$FOXL1_{CTERM}$—6HIS and $FOXL1_{CTERM}$—6HIS remained bound to the Ni column until eluted with higher concentrations of imidazole. Unfortunately, L.5 to L.9 revealed that the expressed protein and target protein co-eluted at similar concentrations of imidazole. However, the major band of the expressed protein (L.6) eluted slightly earlier at 100 mM of imidazole while the major band of $FOXL1_{CTERM}$—6HIS (L.8) eluted later with 200 mM of imidazole.

Three purification experiments were attempted to separate the expressed and target protein based on their slight Ni binding differences. First, a Ni column was run using a higher loading imidazole concentration of 20 mM instead of 5 mM of imidazole as detailed by scheme 5 in Table 2.1. The idea was that a higher imidazole loading concentration could prevent the initial binding of one of the two proteins to the Ni column, such that one protein is removed in the flow-through while the other binds to the Ni column. However, as shown in Figure 3.19, SN—$FOXL1_{CTERM}$—6HIS and $FOXL1_{CTERM}$—6HIS both were able to bind to the Ni column and again co-eluted between 100-200 mM of imidazole. The second purification attempt involved using higher volumes of imidazole washes, as detailed by scheme 6 in Table 2.1, in order to potentially elute the majority of expressed protein before the target protein eluted. As revealed in Figure 3.20, SN—$FOXL1_{MUT}$—6HIS and $FOXL1_{MUT}$—6HIS protein still had significant overlap and co-eluted between 100-200 mM of imidazole. A final Ni affinity purification was tried where more incremental imidazole concentration washes were employed as shown by scheme 7 in Table 2.1. As revealed in Figure 3.21, the major band of SN—$FOXL1_{CTERM}$—6HIS eluted between 80 - 90 mM of imidazole while $FOXL1_{CTERM}$—6HIS eluted at 90 mM imidazole, with significant overlap of the protein. In summary, although Ni affinity chromatography successfully separated the target from the SN fusion protein, it failed to isolate the target protein from the expressed protein.

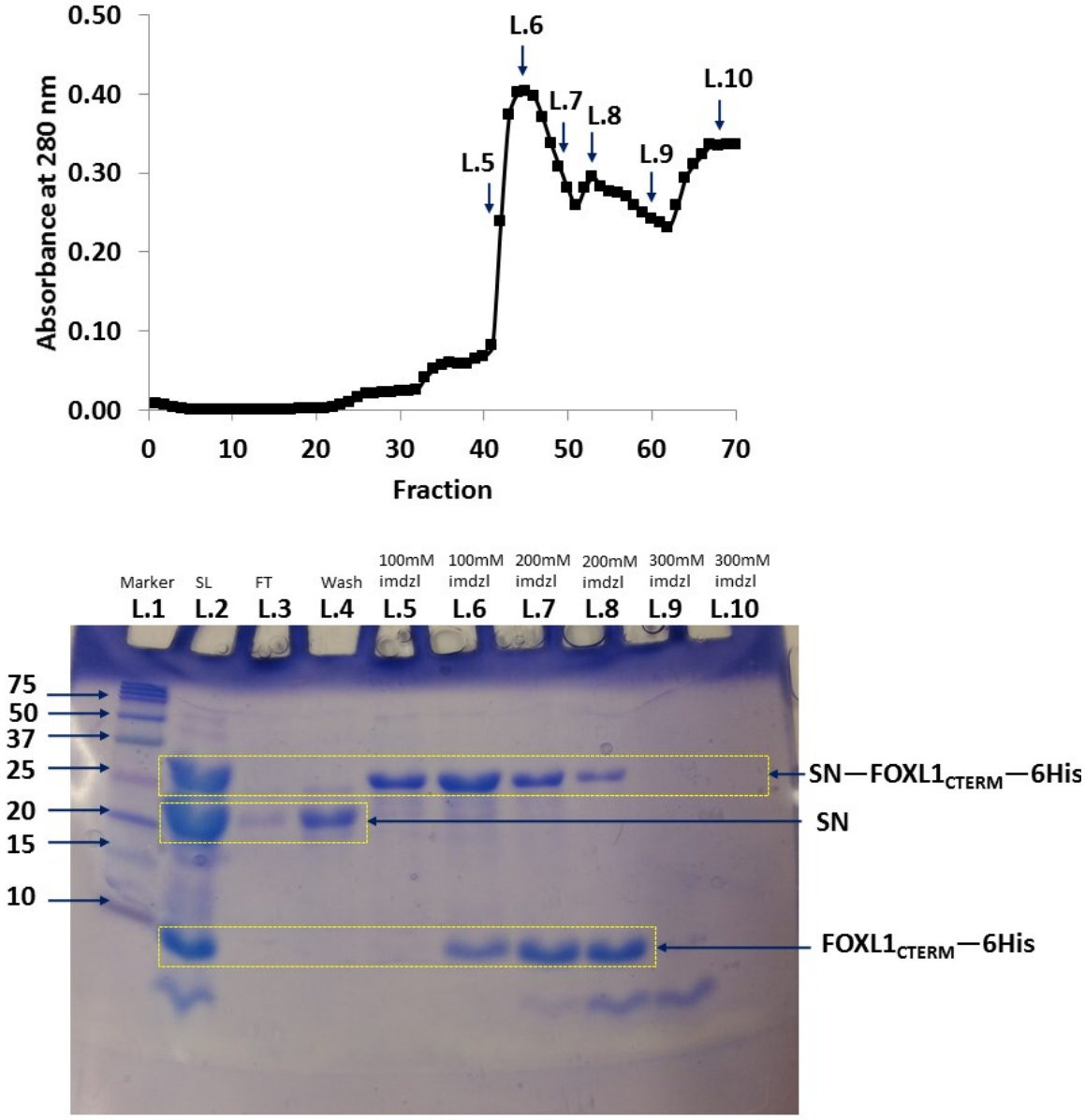


Figure 3.18: UV-vis absorbance (top) and 16.5% tris-tricine Coomassie stained gel (bottom) of the fractions from a second Ni affinity column purification after CNBr digest of SN—FOXL1$_{CTERM}$—6HIS. The elution buffer is shown in scheme 2 of Table 2.1. The SN-fusion protein eluted in the flow-through (L.3) and wash (L.4), while SN—FOXL1$_{CTERM}$—6HIS and FOXL1$_{CTERM}$—6HIS co-eluted between 100-200 mM of imidazole (L.5 - L.9).

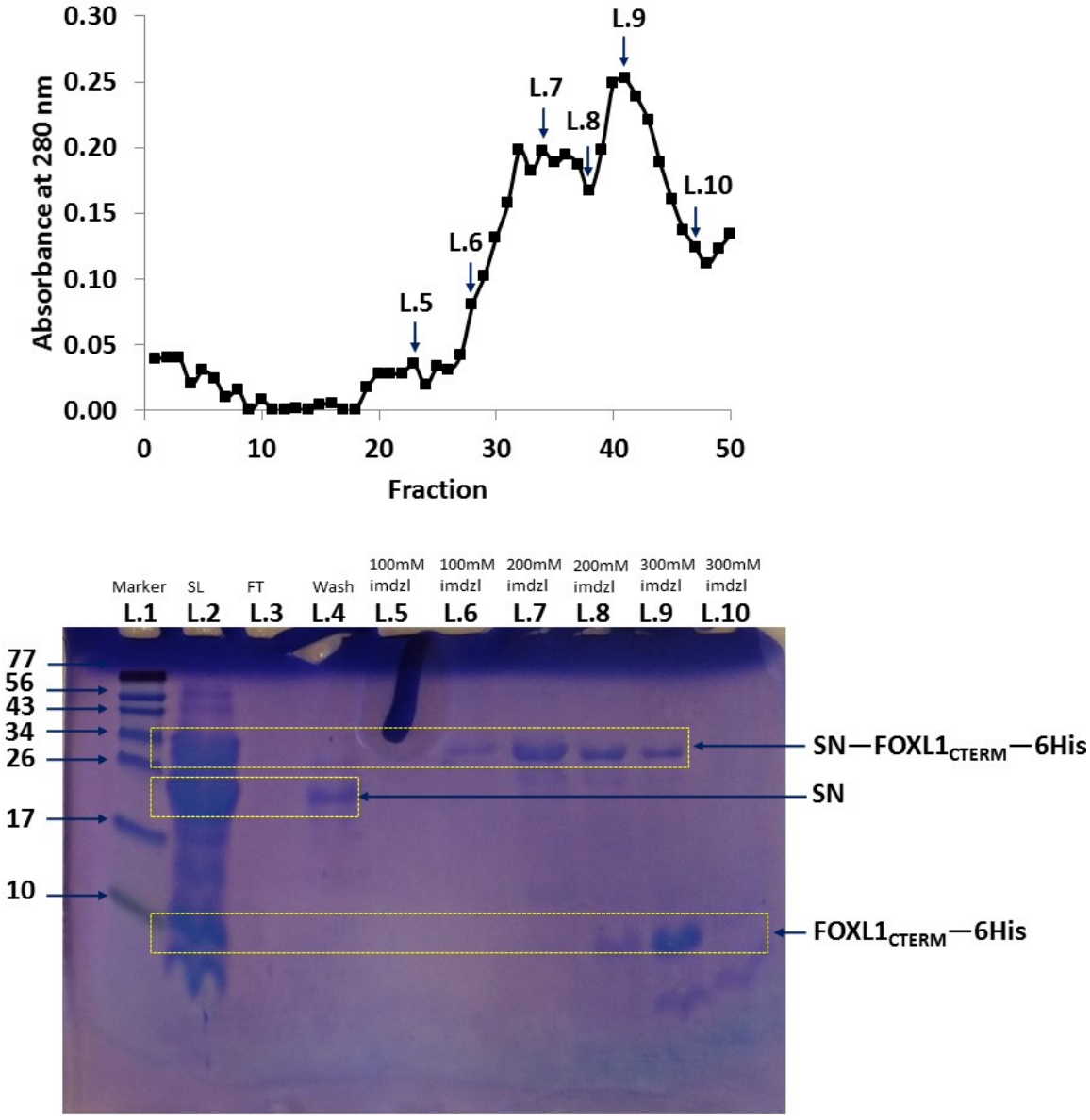


Figure 3.19: UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) following a Ni affinity column purification after CNBr digest of SN—FOXL1$_{CTERM}$—6HIS. A higher initial loading imidazole concentration of 20 mM instead of 5 mM was employed. The elution buffer is shown in scheme 5 of Table 2.1. The SN-fusion protein eluted in the flow-through (L.3) and wash (L.4), while SN—FOXL1$_{CTERM}$—6HIS and FOXL1$_{CTERM}$—6HIS co-eluted between 100-200 mM of imidazole (L.5 - L.9). A higher imidazole loading concentration did not prevent the initial binding of either SN—FOXL1$_{CTERM}$—6HIS or FOXL1$_{CTERM}$—6HIS to the Ni column.

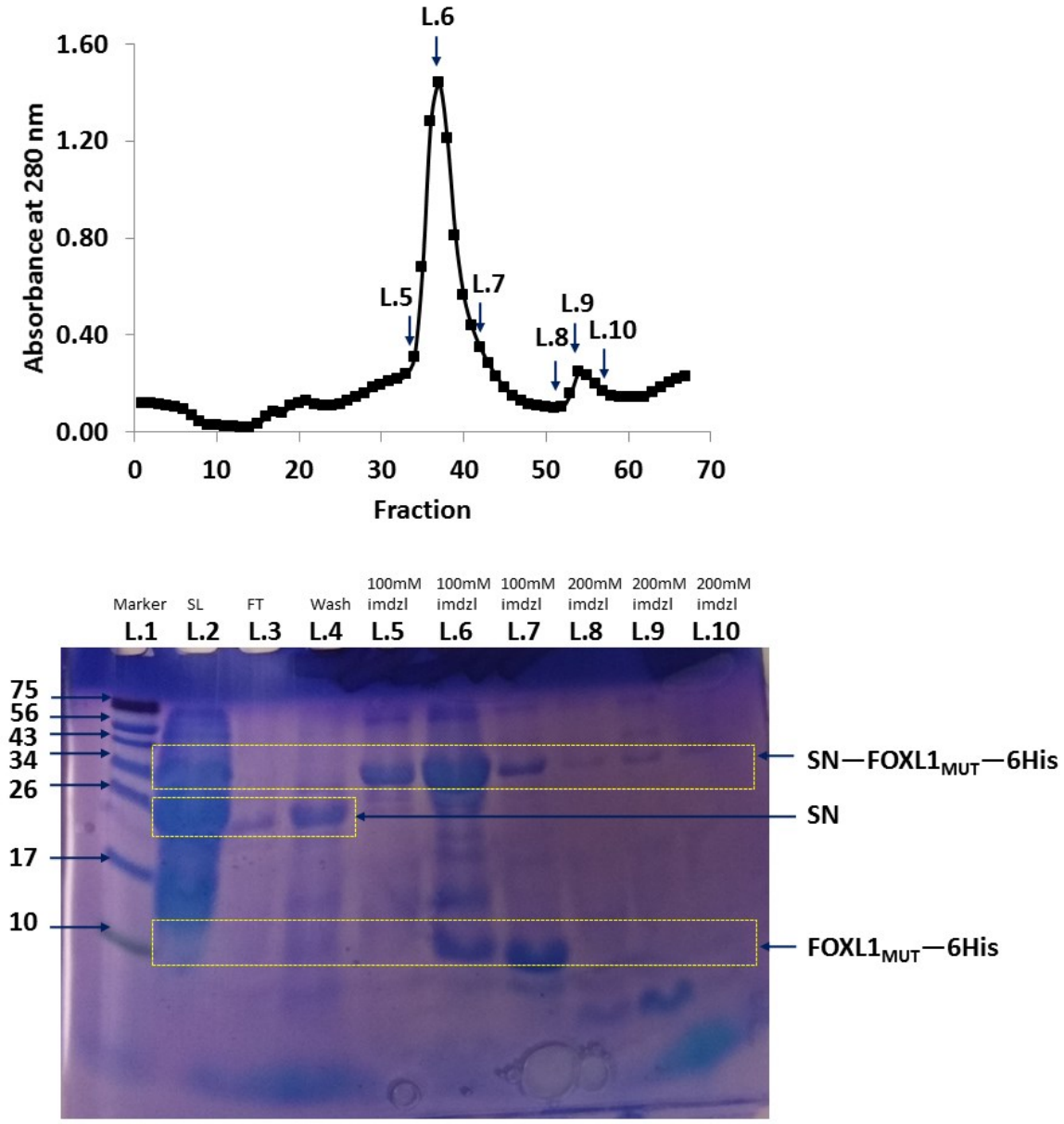


Figure 3.20: UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) of the fractions following a Ni affinity column purification after CNBr digest of SN—FOXL1$_{MUT}$—6HIS using larger volumes of imidazole washes. The elution buffer is shown in scheme 6 of Table 2.1. Even with larger volume of the 50 and 100 mM imidazole wash, the SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS protein had significant overlap and co-eluted between 100-200 mM of imidazole (L.5-L.8).

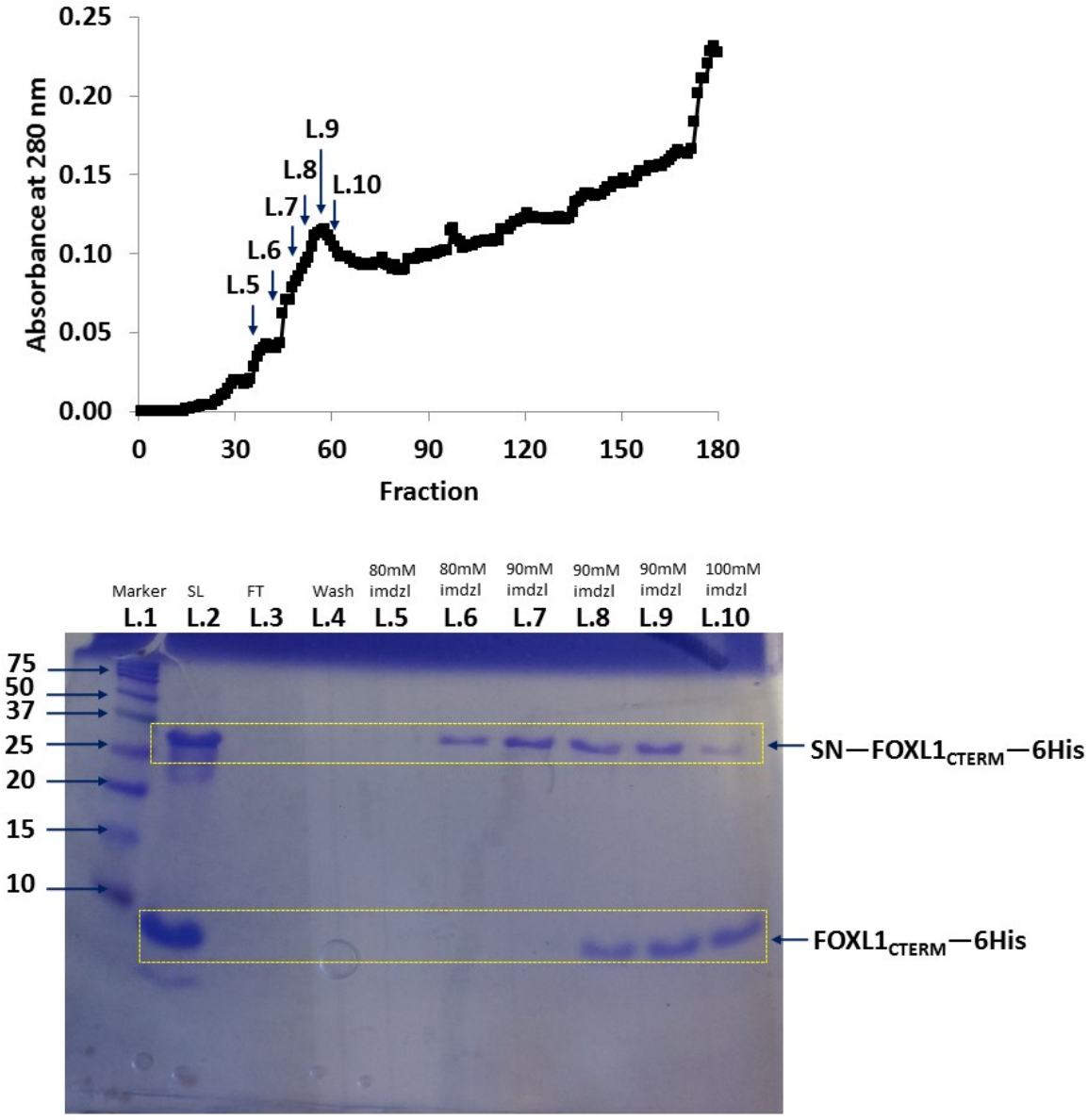


Figure 3.21 UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) of the fractions following a Ni affinity column purification to separate SN—FOXL1$_{CTERM}$—6HIS and FOXL1$_{CTERM}$—6HIS. The elution buffer is shown in scheme 7 of Table 2.1. SN—FOXL1$_{CTERM}$—6HIS eluted between 80 - 90 mM of imidazole (L.6 – L.10) while FOXL1$_{CTERM}$—6HIS eluted at 90 mM imidazole (L.8-L.10) with significant overlap of bands.

Two purification experiments that were attempted to separate the expressed and target protein based on their different molecular weights were microfiltration and size exclusion gel chromatography. First, microfiltration was attempted using a 20 kDa molecular weight cut-off (MWCO) Amacon tube to try and separate SN—FOXL1$_{CTERM}$—6HIS (26.0 kDa) from FOXL1$_{CTERM}$—6HIS (8.1 kDa). The objective was that the lower molecular weight FOXL1$_{CTERM}$—6HIS protein could pass through the porous membrane to the filtrate while the higher molecular weight SN—FOXL1$_{CTERM}$—6HIS protein would remain in the residue. Figure 3.22 showed that microfiltration was not successful because both proteins remained in the residue. It was possible that the urea denaturant in the solvent caused the FOXL1$_{CTERM}$—6HIS protein to become extended (rod shaped), giving it the appearance of a larger size, preventing its passage through the porous membrane.

The second purification tool that was employed to separate the target and expressed protein was size exclusion gel filtration (see Section 2.2.2.13 for details). Three size exclusion column experiments were run in order to optimize the loading concentration and flow rate. The sample preparations are detailed in Table 3.2. In the first experiment, 4 mL of a highly concentrated sample containing SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS (No.1 in Table 3.2) was run through a size exclusion column at a rate of 0.25 mL/min. As seen in Figure 3.23, the major bands from SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS were within 13 mL of each other, which is both the observed and predicted result. Thus, a lower loading concentration was required to separate these proteins. In the second experiment, 1.5 mL of a more dilute sample of FOXL1$_{MUT}$—6HIS and SN—FOXL1$_{MUT}$—6HIS (No.2 in Table 3.2) was run through a size exclusion column at a rate of 0.25 mL/min. As seen in Figure 3.24, size exclusion gel filtration successfully separated the dilute sample of FOXL1$_{MUT}$—6HIS and SN—FOXL1$_{MUT}$—6HIS into completely independent bands. However, the lower concentration made the protein detection by UV-vis impossible and showed only very faint bands in the Coomassie stained gel. Purification with these dilute concentrations would have unfortunately required numerous size exclusion runs to obtain enough target protein for structural studies. Thus, in the final experiment, 1.3 mL of a slightly more concentrated sample of FOXL1$_{MUT}$—6HIS and SN—FOXL1$_{MUT}$—6HIS (No. 3 in Table 3.2) was run through a size exclusion column at a slower rate of 0.15 mL/min. Figure 3.25 revealed that size exclusion gel filtration mostly separated a sample of FOXL1$_{MUT}$—6HIS and SN—FOXL1$_{MUT}$—6HIS with high recovery of

the target protein. Ultimately, although the size exclusion gel filtration purification technique had some overlap of target and expressed protein during elution, it was by far the best method tested that yielded the largest amounts of enriched target protein. For later reference, a sample containing SN—FOXL1$_{CTERM}$—6HIS and FOXL1$_{CTERM}$—6HIS was also successfully separated using size exclusion gel filtration (No. 4 in Table 3.2) as seen in Figure 3.26.

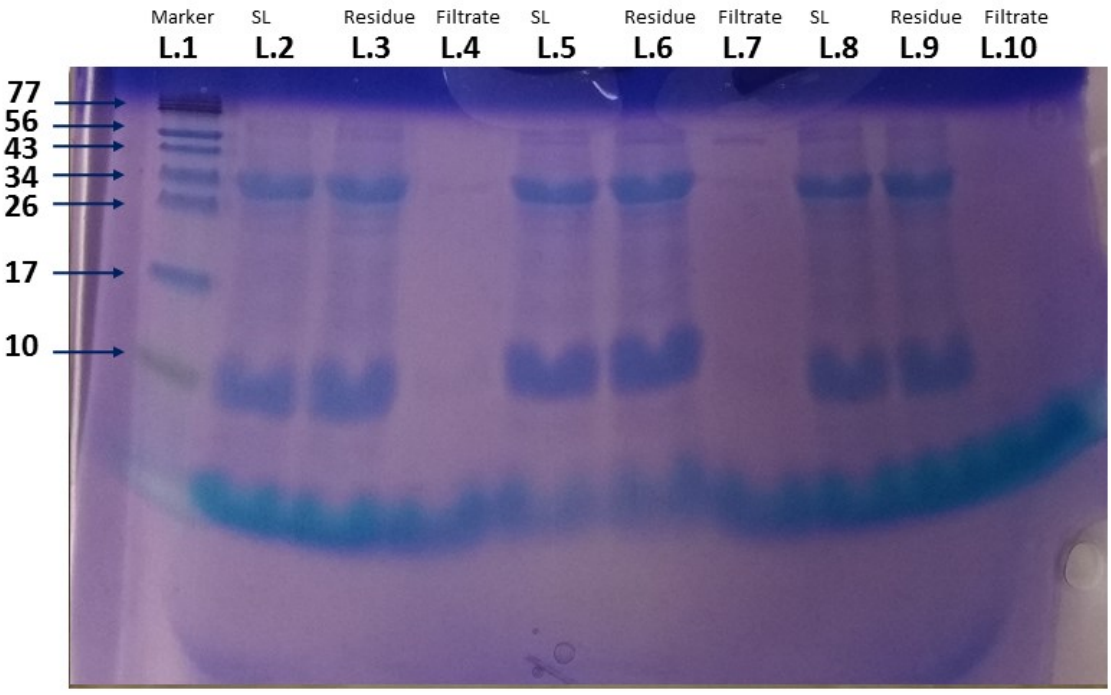


Figure 3.22: Coomassie stained 16.5% tris-tricine gel revealed that microfiltration using a 20 kDa MWCO Amacon tube could not separate FOXL1$_{MUT}$-6HIS and SN-FOXL1$_{MUT}$-6HIS protein. L.2, L.5, and L.8 represent the sample loaded (representing 5mL of combined fraction 36-48 from Figure 3.20); L.3, L.6, and L.9 represent the residue after sample volume was decreased by half; and L.4, L.7, and L.10 was the filtrate which contained small amount of both proteins.

Table 3.2: Size exclusion column details.

No.	Volume Loaded (mL)	Flow rate (mL/min)	Sample preparation
1	4.0	0.25	$FOXL1_{MUT}$—6HIS: Fractions 36-48 (L.6 – L.8 of Figure 3.20) from the Ni affinity column filtration was combined and concentrated from 12.5 mL to 4 mL using a 5000 Da MWCO microfiltration apparatus and then loaded onto size exclusion column.
2	1.5	0.25	$FOXL1_{MUT}$—6HIS: Sample was fraction 47 (L.6 of Figure 3.23c) of the first SEC experiment No. 1
3	1.3	0.15	$FOXL1_{MUT}$—6HIS: Fractions 48-53 (L.16-L.19 of Figure 3.23d) from the first SEC experiment No.1 was concentrated from 9.5 mL to 1.3 mL using a 5000 Da MWCO microfiltration apparatus.
4	1.5	0.17	$FOXL1_{CTERM}$—6HIS: Fraction 38-50 (L.8-L.10 of Figure 3.19) were dialyzed against water using 1 kDa MWCO tubing, freeze dried, dissolved in 1.5 mL of 6 M urea, 0.2% CHAPS, TBS.

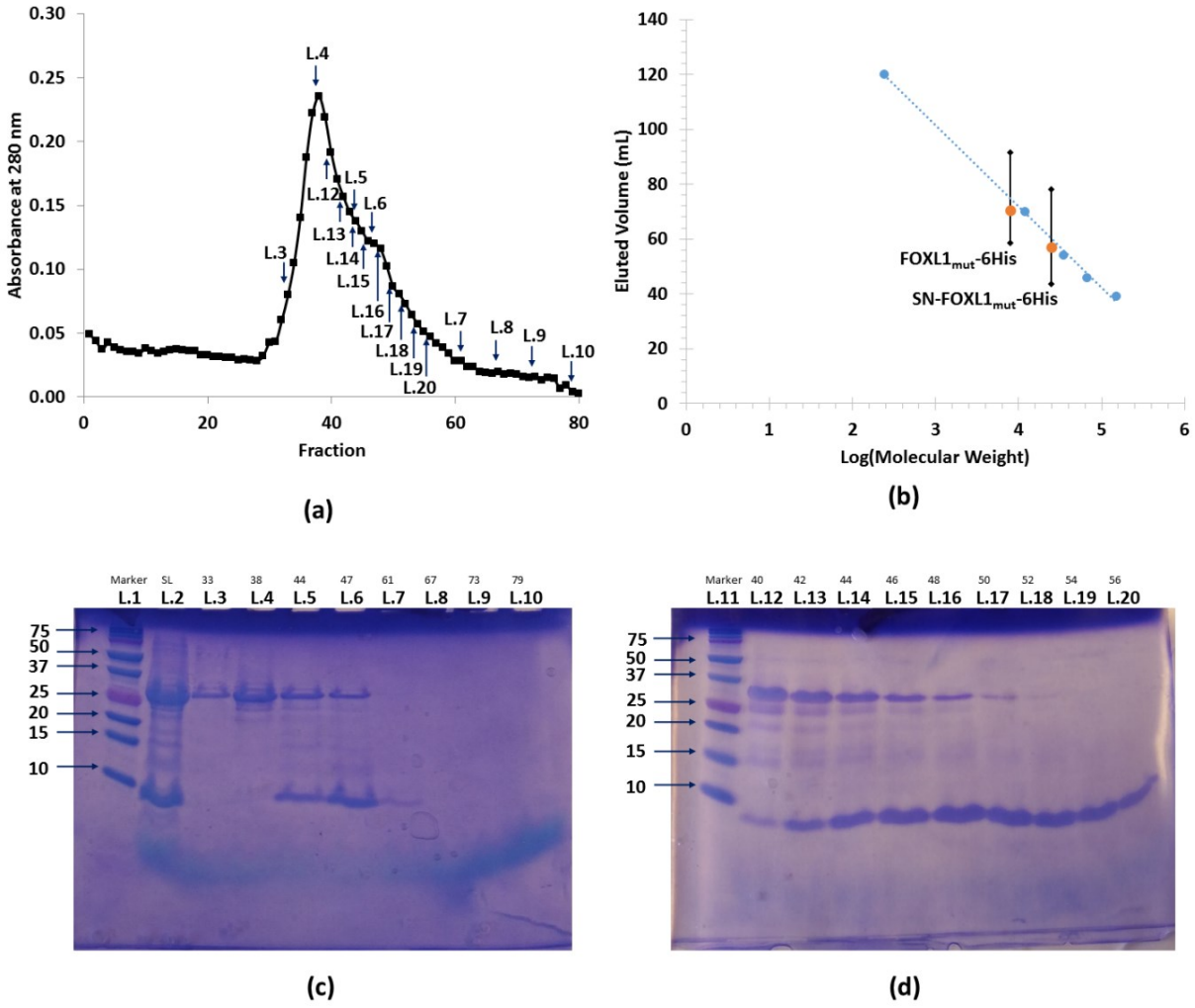


Figure 3.23: Size exclusion column purification to separate $FOXL1_{MUT}$—6HIS and SN—$FOXL1_{MUT}$—6HIS protein done in 6 M urea, 0.2% CHAPS, TBS solvent where 80×1.5mL fractions were collected at a rate of 0.25 mL/min (No. 1 in Table 3.2). (a) UV-vis absorbance of the collected size exclusion fractions shows one major elution peak centered at fraction 44 with a slight shoulder at fraction 61. (b) A plot illustrating the linear relation between the logarithm of the molecular weight and the elution volumes for references molecules, shown in blue. The range of volumes required to elute $FOXL1_{MUT}$—6HIS and SN—$FOXL1_{MUT}$—6HIS are indicated by black lines with an orange dot centered on the fraction containing the majority of each protein. This plot revealed that the major bands of $FOXL1_{MUT}$—6HIS and SN—$FOXL1_{MUT}$—6HIS eluted within 13 mL of each other (both predicted and experimentally observed). (c), (d) Coomassie stained 16.5% tris-tricine gel of the size exclusion column fractions revealed that $FOXL1_{MUT}$—6HIS and SN—$FOXL1_{MUT}$—6HIS protein elute in many of the same fraction as seen in L.5, L.6, and L.12-L.18.

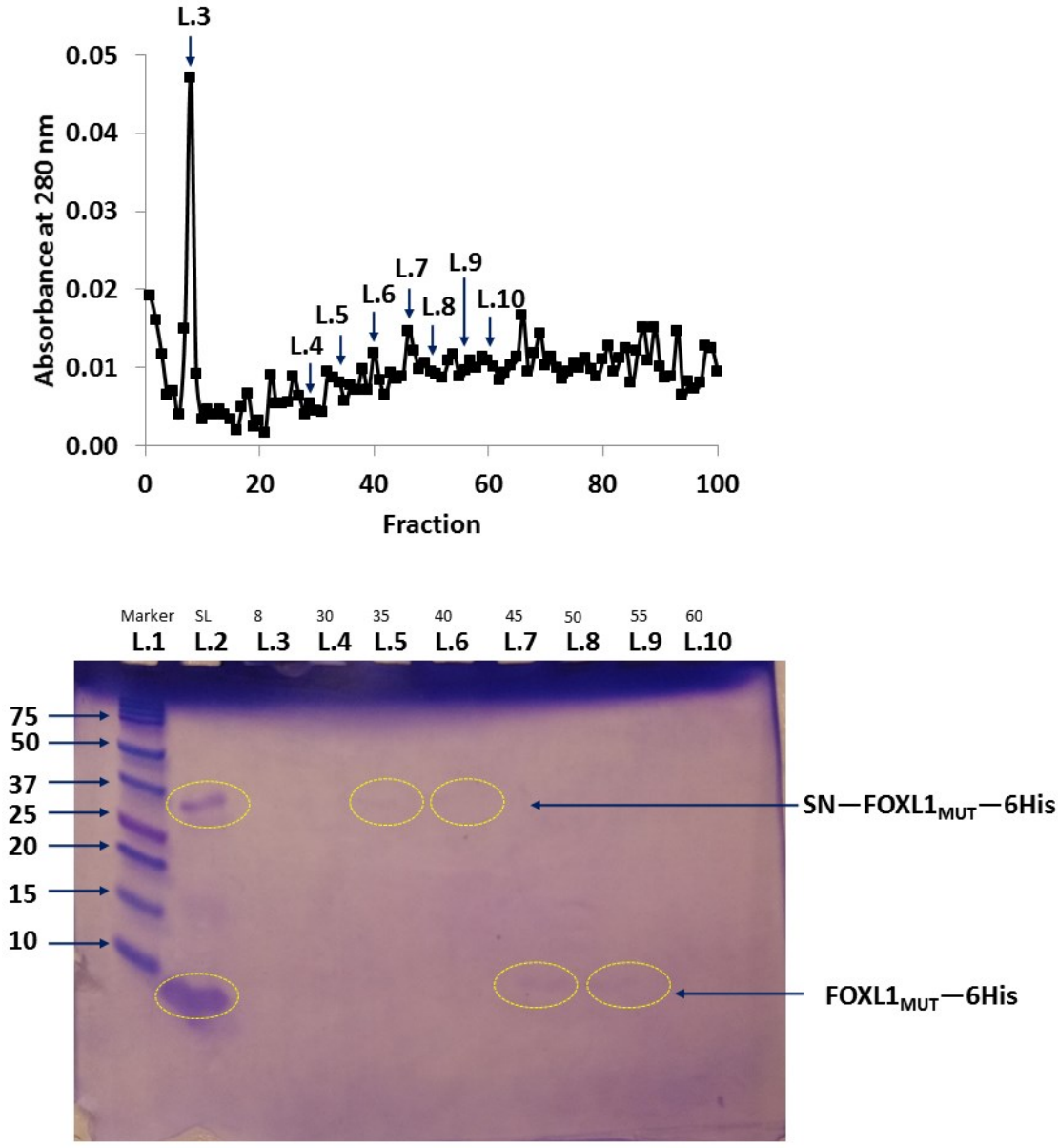


Figure 3.24: UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) of fractions following a size exclusion column purification of a dilute sample containing SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS (L.2). This was done in 6 M urea, 0.2% CHAPS, TBS solvent where 100×1.5mL fractions were collected at a rate of 0.25 mL/min (No. 2 in Table 3.2). UV-vis absorbance of the collected fractions did not show any major peaks. Gel of the size exclusion column fractions showed faint bands of SN-FOXL1$_{MUT}$-6HIS protein in fraction 35 and 40 (L.5 and L.6) and faint bands representing FOXL1$_{MUT}$—6HIS in fractions 50 and 55 (L.8 and L.9.) The gel revealed that size exclusion gel filtration could successfully separate a dilute sample of SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS.

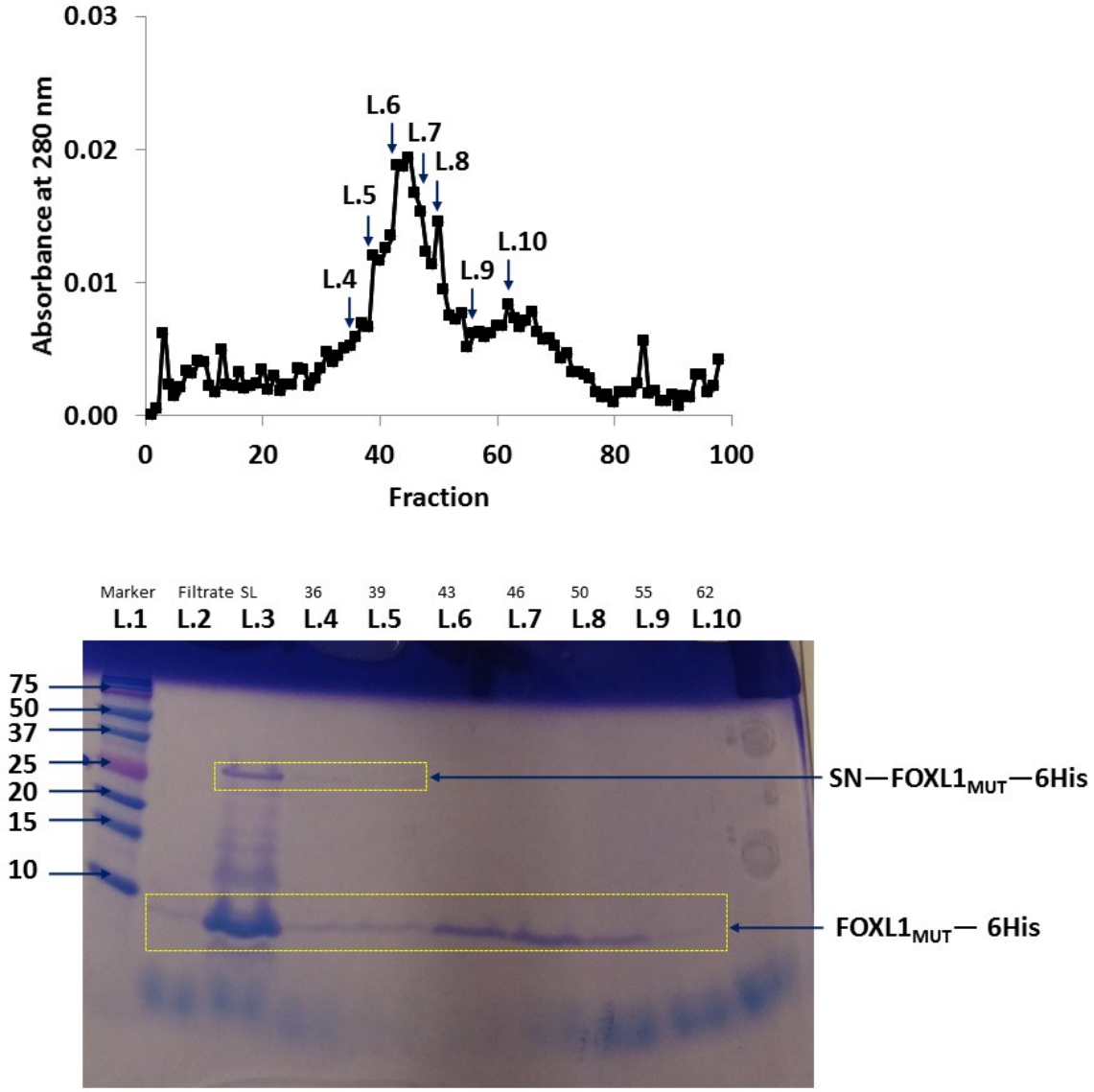


Figure 3.25: UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) of fractions following a size exclusion column purification of a sample containing SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS (L.3). This was done in 6 M urea, 0.2% CHAPS, TBS solvent where 100×1.5mL fractions were collected at a rate of 0.15 mL/min (No. 3 in Table 3.2). (Top) UV-vis absorbance of the collected size exclusion fractions shows a major elution peak centered at fraction 43 (L.6) with a slight shoulder at fraction 39 (L.5) as well as a smaller peak centered at fraction 62 (L.10). (Bottom) Gel of the size exclusion column fractions showed bands representing FOXL1$_{MUT}$—6HIS protein in fractions 35 to 55 (L.4-L.10) and faint bands of SN—FOXL1$_{MUT}$—6HIS in fractions 35-39 (L.4 and L.5). The gel revealed that size exclusion gel filtration could successfully separate a sample of SN—FOXL1$_{MUT}$—6HIS and FOXL1$_{MUT}$—6HIS with high recovery of the target protein.

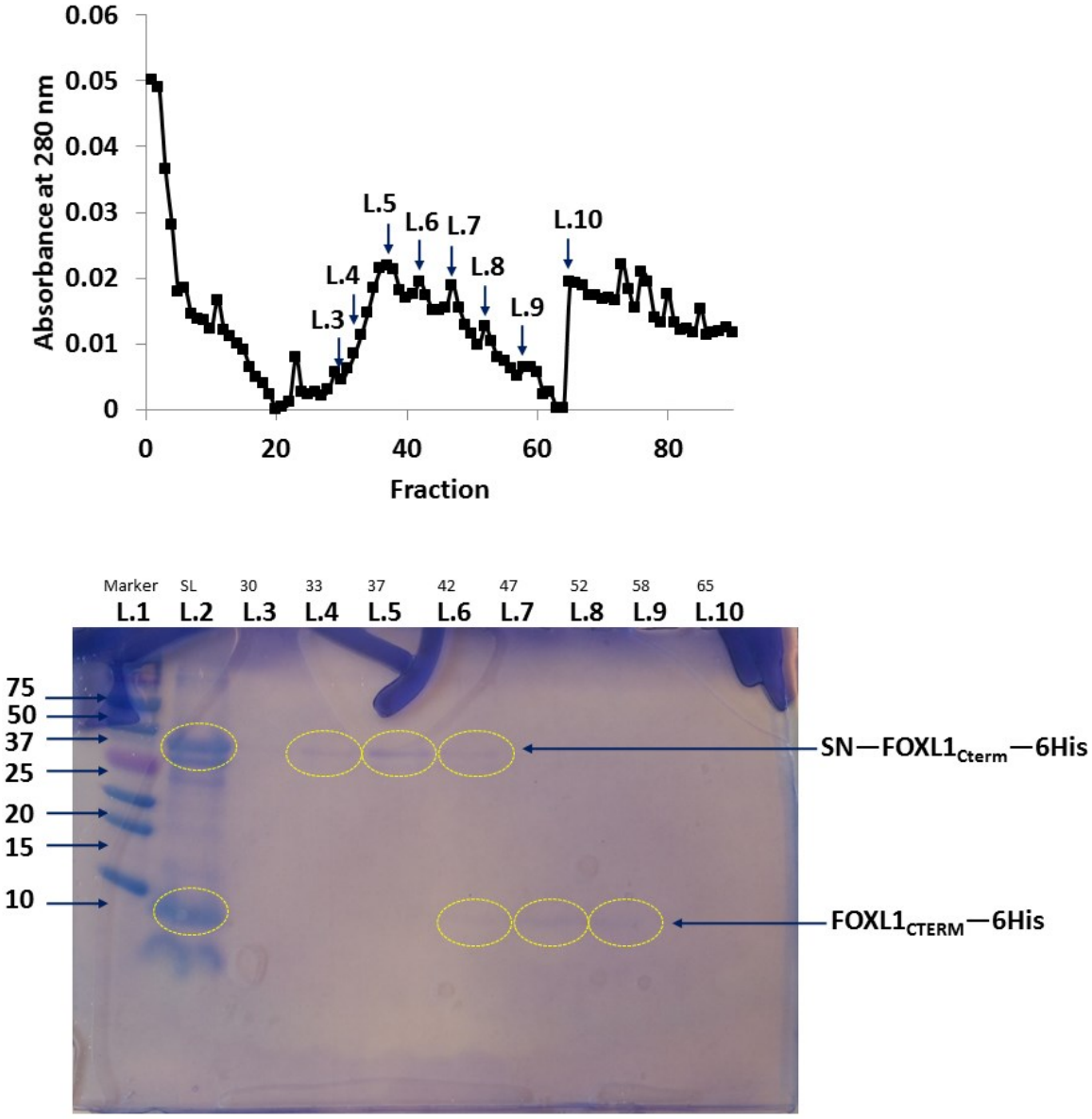


Figure 3.26: UV-vis absorbance (top) and Coomassie stained 16.5% tris-tricine gel (bottom) of fractions following a size exclusion column purification of a sample containing SN—FOXL1$_{CTERM}$—6HIS and FOXL1$_{CTERM}$—6HIS (L.2). This was done in 6 M urea, 0.2% CHAPS, TBS solvent where 90×1.5mL fractions were collected at a rate of 0.17 mL/min (No. 4 in Table 3.2). (Top) UV-vis absorbance of the collected size exclusion fractions shows a major elution peak centered at fraction 37 (L.5). (Bottom) Gel of the size exclusion column fractions showed bands representing FOXL1$_{CTERM}$-6HIS protein in fractions 42 to 58 (L.6-L.10) and faint bands of SN—FOXL1$_{CTERM}$—6HIS in fractions 33-42 (L.4-L.6) The gel reveals that size exclusion gel filtration can successfully separate a sample of FOXL1$_{CTERM}$-6HIS and SN-FOXL1$_{CTERM}$-6HIS.

3.3.4.2. Structural Analysis

After successful enrichment by size exclusion gel filtration, $FOXL1_{CTERM}$—6HIS and $FOXL1_{MUT}$—6HIS were prepared for structural determination by circular dichroism.

Two $FOXL1_{MUT}$—6HIS samples were prepared. The first sample originated from the size exclusion column represented by L.7 to L.9 of Figure 3.24. These fractions were combined, dialyzed against distilled water, and freeze dried. The remaining solid was dissolved in 300 µL of 10 mM potassium phosphate at pH 7. CD was run on this sample to yield the spectrum showed in the top image of Figure 3.27. This spectrum shows a slight minimum, or perhaps an inflection point, around 222 nm and a two-trough minimum at 199 nm and 203 nm. Thus, the protein does not appear to be purely α-helical, β-sheet, or random coil in structure. Instead, this CD spectrum most likely results from the superposition of at least two types of secondary structure. The combination of a random coiled protein spectrum that has negative minima around 200 nm with that of a helical protein spectrum which has negative minima at 208 nm and 212 nm could potentially reproduce the resultant CD spectrum. Thus, it was possible that $FOXL1_{MUT}$—6HIS had a randomly coiled structure with some helical regions.

The second $FOXL1_{MUT}$—6HIS sample that was prepared for circular dichroism had originated from a size exclusion column represented by L.6 to L.9 of Figure 3.25. These fraction were combined, dissolved in 5% acetic acid, dialysed against 2 L of 5% acetic acid, dialyzed against distilled water, and then freeze dried. The remaining solid was dissolved in 250 µL of 10 mM potassium phosphate at pH 7. CD was run on this sample to yield the spectrum showed in the bottom image of Figure 3.27. This CD spectrum also showed the superimposed features of a helical protein (minima at 208 nm and 212 nm) with a random coiled protein (minima around 200 nm).

The $FOXL1_{CTERM}$—6HIS sample that was prepared for circular dichroism came from the combined fractions of a size exclusion column represented by L.7 – L.9 of Figure 3.26. The sample was run through a PD-10 desalting column to replace that the salt and imidazole and with 10 mM potassium phosphate at pH 7. The absorbance of this sample was 0.311 at 280 nm which equated to 50.6 µM of protein. The circular dichroism spectrum of the eluted sample, shown in Figure 3.28, exhibited minima at 208 and 222 nm which suggested a dominant helical shape for

$FOXL1_{CTERM}$—6HIS. However, there is significant noise in the spectra and this limits interpretation. Consequently, the percentage of α-helical character could not be quantified from the spectra.

The final construct, SN—$FOXL1_{CTERM/MUT}$—6HIS, was the most successful of all constructs investigated. First, the expressed protein had high yields and was easily identified via western blot. Upon digest to cleave off the SN-fusion protein, a mixture of expressed, SN-fusion, and the target protein remained. The target protein was successfully enriched, using Ni affinity chromatography to separate out the SN-fusion protein and size exclusion gel filtration to separate out the express protein. Finally, circular dichroism of the two proteins yield distinct differences between $FOXL1_{CTERM}$—6HIS and $FOXL1_{MUT}$—6HIS. While $FOXL1_{CTERM}$—6HIS gave a predominately helical structure, $FOXL1_{MUT}$—6HIS appeared to have a mixed random coil and helical secondary structure.

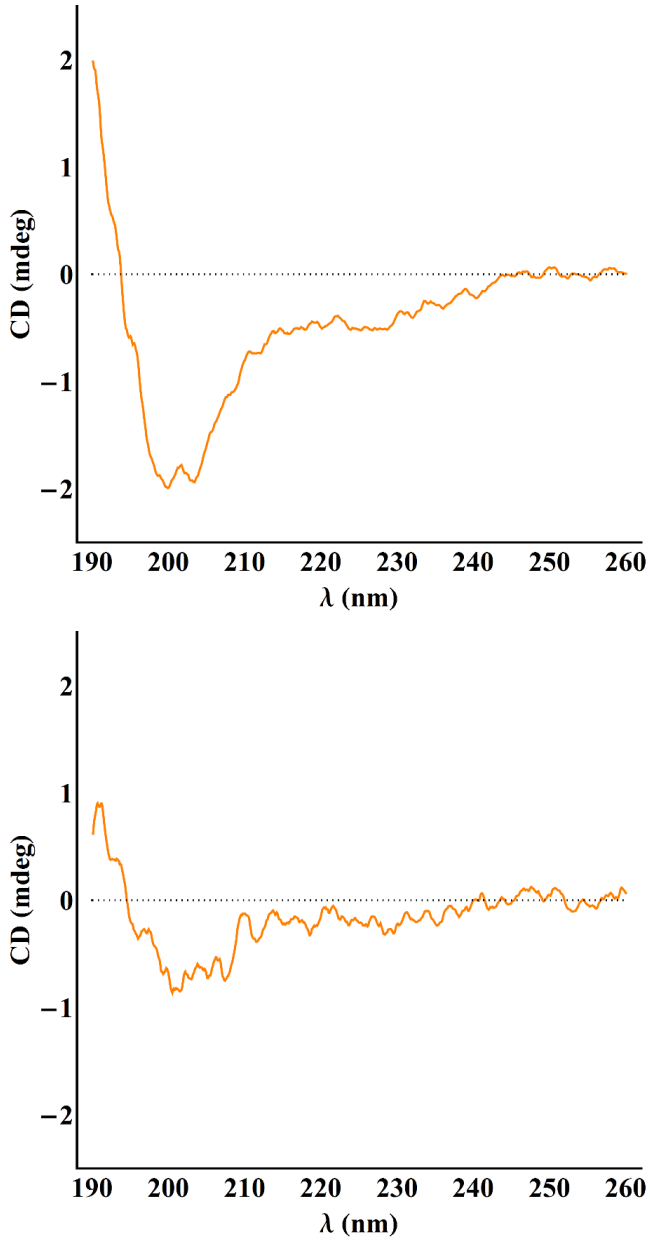


Figure 3.27: Far UV CD spectra of FOXL1$_{MUT}$—6HIS in 10mM of Potassium phosphate at pH 7.0 with a 0.5 mm quartz cuvette at room temperature, where 5 scans were averaged, where purified fractions were obtained from (top) the second size exclusion column (No. 2 in Table 3.2) and (bottom) the third size exclusion column (No. 3 in Table 3.2). Both CD spectra showed the superposition of a helical protein (minima at 208 nm and 212 nm) with a randomly coiled protein (minima between 190-200 nm).

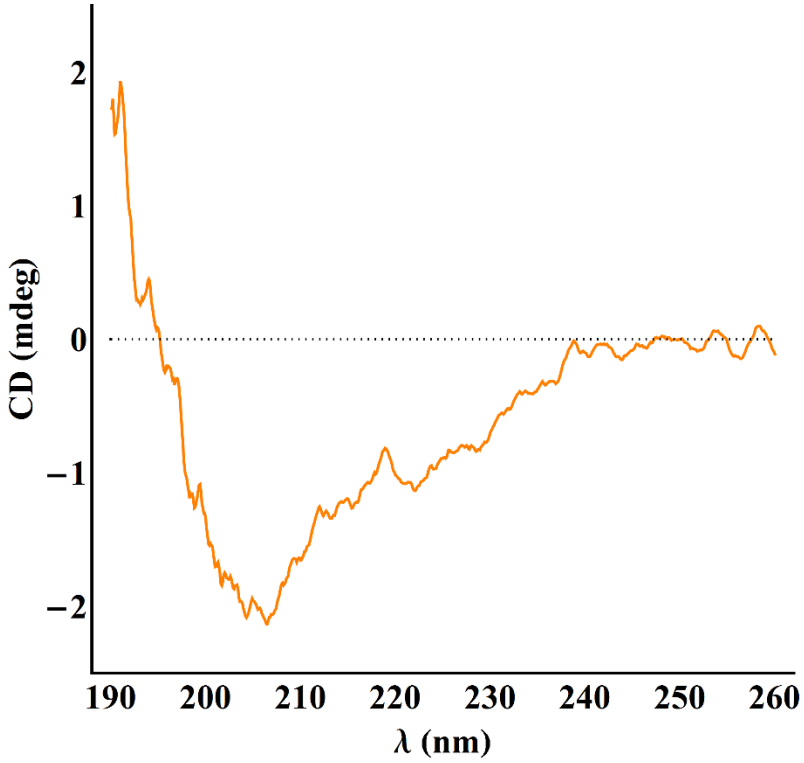


Figure 3.28: Far UV CD spectra of FOXL1$_{CTERM}$—6HIS (50.6 µM) in 10 mM of potassium phosphate at pH 7.0 with a 0.5 mm quartz cuvette at room temperature, where 5 scans were averaged. The CD spectrum showed the features of a helical protein (minima at 208 nm and 212 nm).

3.4. Comparison

There is some agreement between the experimental CD spectra and the computationally determined structures obtained by clustering analysis, which is depicted in Figure 3.29. For $FOXL1_{CTERM}$, the second, third, and fourth most probable structures show a predominantly helical structure that could account for the corresponding CD spectrum. Similarly, for $FOXL1_{MUT}$, the second, third, and fourth most probable structures show a mixed random coil and helical structure that could account for the CD spectrum obtained.

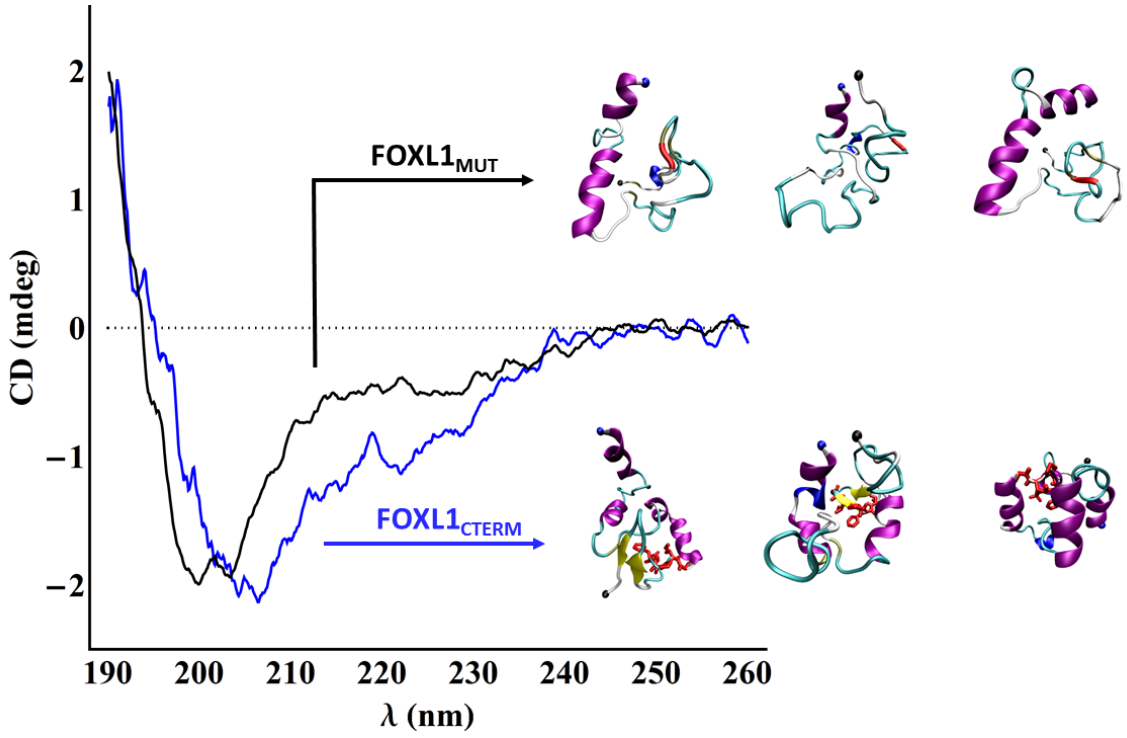


Figure 3.29: Comparison of the experimental CD spectra for $FOXL1_{CTERM}$ and $FOXL1_{MUT}$. The $FOXL1_{MUT}$ spectrum is consistent with a superposition of α-helical and random coiled structures, while the $FOXL1_{CTERM}$ is consistent with an α-helical structure. Clusters extracted from the REMD simulations that are consistent with the CD results are shown.

Protein structures are increasingly being determined using integrative structural biology approaches, where experimental data is combined with computational models. Methods such as Modeler, Rosetta, MELD, and molecular dynamics simulations can employ experimental parameters as restraints in order to predict a reasonable structure that falls within these restraints.[89] For example, the structure of the 201 amino acid protein ALG13 was determined by combining backbone-only NMR data with Rosetta.[90] Currently in this thesis, computational and experimental approaches were employed independently of each other in order to determine the structure of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$, making the agreement between preliminary computational and experimental data remarkable. To validate these preliminary result, further experimental investigation is required. When greater experimental information becomes available, such as NMR data, an assessment of the quality of these preliminary results can be made.

The computational and experimental results both agree that the deletion of GIPFL results in a structural change of the C-terminal domain. This is an important result because function-changing mutations in proteins do not necessarily change its structure. As a case in point, the protein FOXC1 may or may not undergo a structural change depending on the mutation.[91] FOXC1 is a member of the forkhead box family that plays a role in embryonic and ocular (eye) development. Several instances of single-point mutations (where one amino acid is replaced by another) have been linked to Axenfeld-Rieger (AR) syndrome. AR syndrome is an eye disorder that is characterized by irregularities in the front part of the eye, such as a thin and poorly developed iris or abnormalities of the cornea. In a study done by Walter and coworkers,[91] five single point mutations (P79L, P79T, I91S, I91T, and R127H)[§§] in FOXC1, which had been identified in patients with AR syndrome, were studied in order to observe the effect of these mutations on the structure and function of FOXC1 protein. The result of this study showed that (1) the I91S and I91T mutation generates local structural disruptions, (2) the R127H mutation alters the electrostatic charge of the DNA-binding surface, and (3) the P79L and P79T mutation did not significantly change the structure.

[§§] As an example of the mutation notation, a “P79L” mutation denotes that a “P” (proline) amino acid located at residue “79” of the sequence was mutated to a “L” (lysine).

Chapter 4

Conclusion

4.1. Summary

The GIPFL deletion in the C-terminal domain of FOXL1 protein is linked to a human genetic disease. The goal of this research was to acquire structural information about $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ in order to gain insight into the structural differences that caused improper functioning of $FOXL1_{MUT}$ protein. To accomplish this goal, $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ was investigated using: (1) bioinformatics techniques to predict structure from amino acid sequence, (2) computational simulations to fold these proteins into their native structure through conformational sampling techniques, and finally (3) by expressing and structurally characterizing the $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ protein.

First, bioinformatics was employed to predict the disorder, sequence alignment, and secondary structure of FOXL1. The disorder program predicted that a C-terminal domain existed

for FOXL1 which prompted investigation of the 69 most C-terminal residues of FOXL1 ($FOXL1_{CTERM}$) and its corresponding mutant containing the GIPFL deletion ($FOXL1_{MUT}$). One key insight revealed through bioinformatics was that the GIPFL deletion occurred in the most ordered and evolutionary-conserved portion of the C-terminal domain of FOXL1, suggesting that this mutation could severely affect the structure and function of $FOXL1_{CTERM.}$ Unfortunately, reliable secondary structure information was not obtained for $FOXL1_{CTERM}$ due to the absence of homologous protein with known structure.

Secondly, replica exchange molecular dynamics was employed to predict the most probable structures of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ protein. A variety of possible secondary structures were identified for $FOXL1_{CTERM}$ using clustering analysis, including predominantly α-helical structures and mixed α-helical/β-sheet structures. Similarly, a variety of possible secondary structures were identified for $FOXL1_{MUT}$, all of which included a disordered C-terminal region. The N-terminal region of the $FOXL1_{MUT}$ structures were composed either of α-helical or β-sheet structures. Overall, these computationally predicted clusters showed that $FOXL1_{CTERM}$ was more folded and structured than $FOXL1_{MUT}$. These results suggested that the deletion of GIPFL residues in the $FOXL1_{MUT}$ system disrupted its structure and hydrophobic core, causing the mutant to become predominantly randomly coiled.

Interestingly, $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ are the second and third largest proteins folded using PACE, which have 69 and 64 amino acids, respectively.[24,25,45,53] The largest protein published that was correctly folded from a random coil into its native structure using PACE was α3D, a 73 amino acid designed protein that forms a three-helix bundle.[45,53] As well, $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ are the only structures investigated that (potentially) have a significant percentage of intrinsically disordered regions. $FOXL1_{CTERM}$ is predicted to be ~47% disordered by PONDR-FIT as seen in Figure 3.1. $FOXL1_{MUT}$ is even more randomly coiled than $FOXL1_{CTERM}$ based on experimental CD results, which is in agreement with the computationally obtained results by comparing the top four clusters of the REMD simulations.

Finally, progress was made towards expressing, purifying, and structurally characterizing $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ protein. This was accomplished using the construct, SN—$FOXL1_{CTERM/MUT}$—6HIS, which had high expression yields, was easily identified via western blot, was successfully enriched, and allowed for some structural characterization of the target protein

by circular dichroism. The preliminary circular dichroism results suggest that $FOXL1_{CTERM}$—6HIS had a helical structure while $FOXL1_{MUT}$—6HIS was partially helical with some randomly coiled regions.

The combination of bioinformatics, computation simulations, and preliminary experimental results strongly suggests that the deletion of GIPFL residues in $FOXL1_{MUT}$ negatively affect its structure by disrupting the hydrophobic core to yield a more randomly coiled structure. We theorize that this disordering in $FOXL1_{MUT}$ may alter the protein-protein binding surface required to bind to its co-regulatory protein and thus prevent the mutant FOXL1 protein from functioning correctly.

4.2. Future Work

4.2.1. Structural Studies

The next step in the structural investigation of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ is to employ NMR to elucidate their tertiary structures and identify any disordered regions in these target proteins. NMR is an appropriate technique because (1) the disordered regions in these target proteins (as indicated by CD and PONDR-FIT) will likely make protein crystallization too difficult for use in X-ray crystallography,[63] (2) EM would not be expected to provide atomic resolution data information, and (3) these proteins are too small for EM studies [60] The relatively small molecular weights of the target protein (~8.5 kDa) make these proteins suitable for solution NMR studies, such as 2D TOCSY and 2D NOESY.[80] In many cases, proteins with ambiguous peak assignment cannot be determined exclusively from homonuclear ^{1}H NMR data.[80] In this case, heteronuclear NMR techniques can be explored. This would require developing a procedure for isotopic labelling of the protein with ^{15}N or ^{13}C through expression in minimal labelled media.[92] In this method, cells are first grown in unlabelled rich media and then transferred into a small amount of ^{15}N or ^{13}C upon reaching a high cell density for a short growth period.[92] Then, the cells are induced to produce protein for a short period of time, resulting in isotopically labelled protein.

From the acquired homonuclear and heteronuclear NMR data, the solution structures of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ can potentially be determined. These structures can then be refined by combining the obtained spatial information with molecular dynamics simulations.[76] The atoms

with known contacts can be constrained in in the simulation, and then MD can be performed to obtain an equilibrated structure.

4.2.2. Functional Studies

The long term goal of this research is to acquire a thorough functional study of $FOXL1_{CTERM}$ and $FOXL1_{MUT}$ in order to further understand how this mutation causes a human genetic disease. $FOXL1_{CTERM}$ would be studied first in order to (1) determine the co-regulatory protein(s) that can bind to the wildtype, (2) identify the key amino acid residues that constitute the binding site, and (3) determine the structure and the binding affinities of the resultant protein–$FOXL1_{CTERM}$ complexes. Equipped with this information, the mutant protein can be investigated to (4) determine if the mutant is also capable of binding the same co-regulatory proteins as the wildtype and, if so, (5) investigate the effect of the mutation on the binding sites and binding affinities.

In order to perform functional analysis experiments, the wildtype and mutant $FOXL1_{CTERM}$ protein must first be expressed and enriched, which can be accomplished using the methodology detailed in this thesis. Pull-down assays can then be employed to screen for putative protein binding partners of $FOXL1_{CTERM}$.[93] In a pull down assay, the target protein is tagged and captured on an immobilized affinity column that binds to the tag. Then, a protein source (e.g. cell lysate) that contains potential binding partners is passed through the affinity column. The non-binding proteins pass through the column and end up in the flow-through and wash. The bound $FOXL1_{CTERM}$ – co-regulatory protein complexes are then eluted from the column using a competing analyte. To determine the identity of the co-regulatory proteins, protein band isolation from a polyacrylamide gel, tryptic digestion of the isolated protein, and mass spectrometric identification of digested peptides can be completed. Once the potential co-regulatory proteins have been determined, multidimensional NMR techniques, such as NOESY and ^{15}N-HSQC,[65,79] can then be employed to identify the key amino acid residues that constitute the binding site of the $FOXL1_{CTERM}$ – co-regulatory protein complexes.[80] Since the chemical shift of a nucleus in NMR is very sensitive to its chemical environment, the binding site and binding interface can often be detected from changes in NMR resonance frequencies that occur when the complex is formed.[65] Computer simulations of the protein-protein complexes within the NMR-derived structure can also be performed to refine the protein structures and to identify the binding sites and interface.

Following this, affinity electrophoresis can used to estimate the binding affinity of the protein-protein interaction.

With respect to the mutant protein, a pull-down assay can also be employed in order to determine if the mutant binds to the same co-regulatory proteins as the wildtype.[93] If the mutant binds to the same co-regulatory proteins then multidimensional NMR experiments, computational simulations, and binding energies will also be determined for the potential mutant – co-regulatory protein complexes. If the mutant does not bind to the same co-regulatory proteins, then this provides a first explanation as to why the mutant FOXL1 protein does not function correctly. Equipped with this information, we can attempt to understand the link between the mutated FOXL1 protein and the human genetic disease.

References

(1) Nelson, D. L., and Cox, M. M. (2008) Lehninger Principles of Biochemistry 5th ed. W. H. Freeman and Company, NY.

(2) Vivekanandan, S., Brender, J. R., Lee, S. Y., and Ramamoorthy, A. (2011) A partially folded structure of amyloid-beta(1-40) in an aqueous environment. *Biochem. Biophys. Res. Commun. 411*, 312–316.

(3) Lehmann, O. J., Sowden, J. C., Carlsson, P., Jordan, T., and Bhattacharya, S. S. (2003) Fox's in development and disease. *Trends Genet. 19*, 339–344.

(4) Benayoun, B. A., Caburet, S., and Veitia, R. A. (2011) Forkhead transcription factors: key players in health and disease. *Trends Genet. 27*, 224–232.

(5) Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein. *Proteins Struct. Funct. Genet. 42*, 38–48.

(6) Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., and Uversky, V. N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta. 1804*, 996–1010.

(7) Fritzenwanker, J. H., Gerhart, J., Freeman, R. M., and Lowe, C. J. (2014) The Fox/Forkhead transcription factor family of the hemichordate Saccoglossus kowalevskii. *EvoDevo. 5*, 17.

(8) Madison, B. B., McKenna, L. B., Dolson, D., Epstein, D. J., and Kaestner, K. H. (2009) FoxF1 and FoxL1 link hedgehog signaling and the control of epithelial proliferation in the developing stomach and intestine. *J. Biol. Chem. 284*, 5936–5944.

(9) Wotton, K. R., and Shimeld, S. M. (2011) Analysis of lamprey clustered Fox genes: Insight into Fox gene evolution and expression in vertebrates. *Gene. 489*, 30–40.

(10) Shin, S., Upadhyay, N., Greenbaum, L. E., and Kaestner, K. H. (2015) Ablation of Foxl1-Cre-labeled hepatic progenitor cells and their descendants impairs recovery from liver injury. *Gastroenterology. 148*, 192–202.

(11) Yang, F.-Q., Yang, F.-P., Li, W., Liu, M., Wang, G.-C., Che, J.-P., Huang, J.-H., and Zheng, J.-H. (2014) Foxl1 inhibits tumor invasion and predicts outcome in human renal cancer. *Int. J. Clin. Exp. Pathol. 7*, 110–122.

(12) Perreault, N., Katz, J. P., Sackett, S. D., and Kaestner, K. H. (2001) Foxl1 controls the Wnt/beta-catenin pathway by modulating the expression of proteoglycans in the gut. *J. Biol. Chem. 276*, 43328–43333.

(13) Qin, Y., Gong, W., Zhang, M., Wang, J., Tang, Z., and Quan, Z. (2014) Forkhead box L1 is frequently downregulated in gallbladder cancer and inhibits cell growth through apoptosis induction by mitochondrial dysfunction. *PloS ONE. 9*, e102084.

(14) Carlsson, P., and Mahlapuu, M. (2002) Forkhead transcription factors: key players in development and metabolism. *Dev. Biol. 250*, 1–23.

(15) Kaestner, K. H., Knöchel, W., and Martínez, D. E. (2000) Unified nomenclature for the winged helix / forkhead transcription factors. *Genes Dev. 14*, 142–146.

(16) Pierrou, S., Hellqvist, M., Samuelsson, L., Enerbäck, S., and Carlsson, P. (1994) Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. *EMBO J. 13*, 5002–5012.

(17) Park, S. E., Oh, K. W., Lee, W. Y., Baek, K. H., Yoon, K. H., Son, H. Y., Lee, W. C., and Kang, M. Il. (2014) Association of osteoporosis susceptibility genes with bone mineral density and bone metabolism related markers in Koreans: The Chungju Metabolic Disease Cohort (CMC) study. *Endocr. J. 61*, 1069–1078.

(18) Zhang, L., Choi, H. J., Estrada, K., Leo, P. J., Li, J., Pei, Y.-F., Zhang, Y., Lin, Y., Shen, H., Liu, Y.-Z., Liu, Y., Zhao, Y., Zhang, J.-G., Tian, Q., Wang, Y., Han, Y., Ran, S., Hai, R., Zhu, X.-Z., Wu, S., Yan, H., Liu, X., Yang, T.-L., Guo, Y., Zhang, F., Guo, Y., Chen, Y., Chen, X., Tan, L., Zhang, L., Deng, F.-Y., Deng, H., Rivadeneira, F., Duncan, E. L., Lee, J. Y., Han, B. G., Cho, N. H., Nicholson, G. C., McCloskey, E., Eastell, R., Prince, R. L., Eisman, J. A., Jones, G., Reid, I. R., Sambrook, P. N., Dennison, E. M., Danoy, P., Yerges-Armstrong, L. M., Streeten, E. A., Hu, T., Xiang, S., Papasian, C. J., Brown, M. A., Shin, C. S., Uitterlinden, A. G., and Deng, H.-W. (2014) Multistage genome-wide association meta-analyses identified two new loci for bone mineral density. *Hum. Mol. Genet. 23*, 1923–1933.

(19) Zufferey, F., Martinet, D., Osterheld, M.-C., Niel-Bütschi, F., Besuchet, S. N., Beckmann, J. S., Xia, Z., Stankiewicz, P., Langston, C., and Fellmann, F. (2013) 16q24.1 microdeletion in a premature newborn: usefulness of array-based comparative genomic hybridization (array CGH) in persistent pulmonary hypertension of the newborn. *Pediatr. Crit. Care. Med. 12*, e427–e432.

(20) Altschup, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol. 215*, 403–410.

(21) Rost, B., and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol. 232*, 584–599.

(22) Wallner, B., and Elofsson, A. (2005) All are not equal: a benchmark of different homology modeling programs. *Protein Sci. 14*, 1315–1327.

(23) John, B., and Sali, A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res. 31*, 3982–3992.

(24) Han, W., Wan, C.-K., Jiang, F., and Wu, Y.-D. (2010) PACE force field for protein simulations. 1. Full parameterization of version 1 and verification. *J. Chem. Theory Comput. 6*, 3373–3389.

(25) Han, W., Wan, C.-K., and Wu, Y.-D. (2010) PACE force field for protein simulations. 2. Folding simulations of peptides. *J. Chem. Theory Comput. 6*, 3390–3402.

(26) Daggett, V. (2002) Molecular dynamics simulations of the protein unfolding/folding reaction. *Acc. Chem. Res. 35*, 422–429.

(27) Zhou, R. (2007) Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol. Biol. 350*, 205–223.

(28) Bank, D. (1996) PHD : Prediction 1D protein structure by profile-based neural networks. *Methods. 266*, 525–539.

(29) Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389–3402.

(30) Kaczanowski, S., and Zielenkiewicz, P. (2010) Why similar protein sequences encode similar three-dimensional structures? *Theor. Chem. Acc. 125*, 643–650.

(31) Schwartz, R. M., and Dayhoff, M. O. (1978) Matrices for detecting distant relationships, in *Atlas of Protein Sequence and Structure* Vol. 5., pp 353–358. National Biomedical Research Foundation, Washington, DC.

(32) Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A. 89*, 10915–10919.

(33) Rost, B., Yachdav, G., and Liu, J. (2004) The PredictProtein server. *Nucleic Acids Res. 32*, W321–W326.

(34) Chu, W.-T., Zhang, J.-L., Zheng, Q.-C., Chen, L., and Zhang, H.-X. (2013) Insights into the folding and unfolding processes of wild-type and mutated SH3 domain by molecular dynamics and replica exchange molecular dynamics simulations. *PloS ONE. 8*, e64886.

(35) Zhang, Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol. 18*, 342–348.

(36) Simons, J. (2003) An Introduction to Theoretical Chemistry. Cambridge University Press, New York.

(37) Frank, J. (2006) Introduction to Computational Chemistry 2nd ed. Wiley, West Sussex.

(38) Karplus, M., and Mccammon, J. A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol. 9*, 646–652.

(39) Ota, K., Makoto, M., Milford, E. L., Mackin, G. A., Weiner, H. L., and Hafler, D. A. (1990) Molecular dynamics simulations in biology. *Lett. to Nat. 346*, 183–187.

(40) Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012) Biomolecular simulation: A computational microscope for molecular biology. *Annu. Rev. Biophys. 41*, 429–452.

(41) Dill, K. A., and Maccallum, J. L. (2012) The protein-folding problem, 50 years on. *Science 338*, 1042–1047.

(42) Guardiani, C., Livi, R., and Cecconi, F. (2010) Coarse grained modeling and approaches to protein folding. *Curr. Bioinform. 5*, 217–240.

(43) Beauchamp, K. A., Lin, Y.-S., Das, R., and Pande, V. S. (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR neasurements. *J. Chem. Theory Comput. 8*, 1409–1414.

(44) Scheraga, H. A., Khalili, M., and Liwo, A. (2007) Protein-folding dynamics: overview of molecular simulation techniques. *Annu. Rev. Phys. Chem. 58*, 57–83.

(45) Han, W., and Schulten, K. (2012) Further optimization of a hybrid united-atom and coarse-grained force field for folding simulations: Improved backbone hydration and interactions between charged side chains. *J. Chem. Theory Comput. 8*, 4413–4424.

(46) Wabik, J., Kmiecik, S., Gront, D., Kouza, M., and Koliński, A. (2013) Combining coarse-grained protein models with replica-exchange all-atom molecular dynamics. *Int. J. Mol. Sci. 14*, 9893–9905.

(47) Han, W., Wan, C., and Wu, Y. (2008) Coarse-grained solvent model: Solvation free energies. *J. Chem. Theory Comput. 4*, 1891–1901.

(48) Marrink, S. J., de Vries, A. H., and Mark, A. E. (2004) Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B 108*, 750–760.

(49) Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B 111*, 7812–7824.

(50) Gu, J., Bai, F., Li, H., and Wang, X. (2012) A generic force field for protein coarse-grained molecular dynamics simulation. *Int. J. Mol. Sci. 13*, 14451–14469.

(51) Shih, A. Y., Freddolino, P. L., Arkhipov, A., and Schulten, K. (2007) Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. *J. Struct. Biol. 157*, 579–592.

(52) Arkhipov, A., Yin, Y., and Schulten, K. (2008) Four-scale description of membrane sculpting by BAR domains. *Biophys. J. 95*, 2806–2821.

(53) Han, W., and Schulten, K. (2013) Characterization of folding mechanisms of Trp-cage and WW-domain by network analysis of simulations with a hybrid-resolution model. *J. Phys. Chem. B. 117*, 13367–13377.

(54) Han, W., and Wu, Y.-D. (2007) Coarse-grained protein model coupled with a coarse-grained water model: Molecular dynamics study of polyalanine-based peptides. *J. Chem. Theory Comput. 3*, 2146–2161.

(55) Heyer, L. J., Kruglyak, S., and Yooseph, S. (1999) Exploring expression data: Identification and analysis of coexpressed genes exploring expression data. *Genome Res.* 1106–1115.

(56) Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F., and Mark, A. E. (1999) Peptide folding: When simulation meets experiment. *Angew. Chemie Int. Ed. 38*, 236–240.

(57) Corrêa, D. H. A., and Ramos, C. H. I. (2009) The use of circular dichroism spectroscopy to study protein folding, form and function. *African J. Biochem. Res. 3*, 164–173.

(58) Greenfield, N. (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc. 1*, 2876–2890.

(59) Hofmann, A. (2009) Principles and Techniques of Biochemistry and Molecular Biology (Wilson, K., and Walker, J., Eds.) 7th ed., pp 511–516. Cambridge University Press, New York.

(60) Boekema, E. J., Folea, M., and Kouřil, R. (2009) Single particle electron microscopy. *Photosynth. Res. 102*, 189–196.

(61) Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. (2013) Cryo-electron microscopy: A primer for the non-microscopist. *FEBS J. 280*, 28–45.

(62) Kourkoutis, L. F., Plitzko, J. M., and Baumeister, W. (2012) Electron microscopy of biological materials at the nanometer scale. *Annu. Rev. Mater. Res. 42*, 33–58.

(63) Ilari, A., and Savino, C. (2008) Protein structure determination by X-ray crystallography. *Methods Mol. Biol. 452*, 63–87.

(64) Wlodawer, A., Minor, W., Dauter, Z., and Jaskolski, M. (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J. 275*, 1–21.

(65) Keeler, J. (2010) Understanding NMR Spectroscopy 2nd ed. Wiley, England.

(66) Jacobsen, N. E. (2007) NMR Spectroscopy Expained. Wiley, New Jersey.

(67) Padilla, A., Vuister, G. W., Boelens, R., Kleywegt, A., Cave, J., Parello, J., and Kaptein, R. (1990) Homonuclear three-dimensional 1H NMR spectroscopy of pike parvalbumin. Comparison of short- and medium-range NOEs from 2D and 3D NMR. *J. Am. Chem. Soc. 112*, 5024–5030.

(68) Young, C. L., Britton, Z. T., and Robinson, A. S. (2012) Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications. *Biotechnol. J. 7*, 620–634.

(69) Gopal, G. J., and Kumar, A. (2013) Strategies for the production of recombinant protein in Escherichia coli. *Protein J. 32*, 419–425.

(70) Boyer, R. (2012) Biochemistry Laboratory: Modern Theory and Techniques 2nd ed. Pearson Education, Inc., Upper Saddle River.

(71) Borgia, J. A., and Fields, G. B. (2000) Chemical synthesis of proteins. *Trends Biotechnol. 18*, 243–251.

(72) Schägger, H. (2006) Tricine-SDS-PAGE. *Nat. Protoc. 1*, 16–22.

(73) Mahmood, T., and Yang, P. C. (2012) Western blot: Technique, theory, and trouble shooting. *N. Am. J. Med. Sci. 4*, 429–434.

(74) Smyth, M. S., and Martin, J. H. (2000) X-Ray crystallography. *Mol. Pathol. 53*, 8–14.

(75) Breg, J. N., Sarda, L., Cozzone, P. J., Rugani, N., Boelens, R., and Kaptein, R. (1995) Solution structure of porcine pancreatic procolipase as determined from 1H homonuclear two-dimensional and three-dimensional NMR. *Eur. J. Biochem. 227*, 663–672.

(76) Fogh, R. H., Ottleben, G., Rüterjans, H., Schnarr, M., Boelens, R., and Kaptein, R. (1994) Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy. *EMBO J. 13*, 3936–3944.

(77) Oschkinat, H., Cieslar, C., Gronenborn, A. M., and Clore, G. M. (1989) Three-dimensional homonuclear Hartmann-Hahn-nuclear overhauser enhancement spectroscopy in H2O and its application to proteins. *J. Magn. Reson. 81*, 212–216.

(78) Wijmenga, S. S., and van Mierlo, C. P. (1991) Three-dimensional correlated NMR study of Megasphaera elsdenii flavodoxin in the oxidized state. *Eur. J. Biochem. 195*, 807–822.

(79) Lian, L., and Roberts, G. (2011) Protein NMR Spectroscopy: Principal Techniques and Applications. Wiley, Malaysia.

(80) Morris, G. A., and Emsley, J. W. (2010) Multidimensional NMR Methods for the Solution State. Wiley, Singapore.

(81) Kumar, A., Ernst, R. R., and Wüthrich, K. (1980) A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. *Biochem. Biophys. Res. Commun. 95*, 1–6.

(82) Bax, A., and Davis, D. G. (1985) MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *J. Magn. Reson. 65*, 355–360.

(83) Wuthrich, K. (1986) NMR of Proteins and Nucleic Acids. Wiley, New York.

(84) Denk, W., Baumann, R., and Wagner, G. (1986) Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines. *J. Magn. Reson. 67*, 386–390.

(85) Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD - Visual Molecular Dynamics. *J. Molec. Graph. 14*, 33–38.

(86) Nosé, S. (1984) A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys. 52*, 255–268.

(87) Hoover, W. G. (1985) Canonical dynamics: equilibrum phase-space distribution. *Phys. Rev. A 31*, 1695–1697.

(88) Duong-Ly, K. C., and Gabelli, S. B. (2014) Troubleshooting recombinant protein expession. *Methods Enzymol. 541*, 209–229.

(89) MacCallum, J. L., Perez, A., and Dill, K. A. (2015) Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *PNAS 112*, 6985–6990.

(90) Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Liu, G., Ramelot, T., Eletsky, A., Szyperski, T., Kennedy, M., Prestegard, J., Montelione, G. T., and Baker, D. (2010) NMR Structure Determination for Larger Proteins Using Backbone-Only Data. *Science 327*, 1014–1018.

(91) Saleem, R. A., Banerjee-Basu, S., Berry, F. B., Baxevanis, A. D., and Walter, M. A. (2003) Structural and functional analyses of disease-causing missense mutations in the forkhead domain of FOXC1. *Hum. Mol. Genet. 12*, 2993–3005.

(92) Bracken, C., Marley, J., and Lu, M. (2001) A method for efficient isotopic labeling of recombinant proteins. *J Biomol NMR 20*, 71–75.

(93) Nguyen, T. N., and Goodrich, J. A. (2006) Protein-protein interaction assays: eliminating false positive interactions. *Nat. Methods 3*, 135–139.